UNIVERSIDAD AUTÓNOMA DE MADRID

Facultad de Ciencias

Departamento de Biología Molecular

Advances in Bioinformatics: Contributions to High-Throughput Proteomics-Based Identification, Quantification and Systems Biology

TESIS DOCTORAL

Marco Trevisan Herraz

Codirectores de tesis:
Prof. Jesús Vázquez Cobos
Dr. Elena Bonzón Kulichenko

Madrid, 2016

A mis padres, Fausto y Dulce María, que vieron el principio

A mis hijos, Leonardo y Matteo, que vieron el final

Y a Marga, que estuvo a mi lado todo este tiempo

Agradecimientos

Sì come per levar, donna, si pone
in pietra alpestra e dura
una viva figura,
che là più cresce u' più la pietra scema;

tal alcun'opre buone,
per l'alma che pur trema,
cela il superchio della propria carne
co' l'inculta sua cruda e dura scorza.

Tu pur dalle mie streme
parti può sol levarne,
ch'in me non è dime voler né forza

Michelangelo, «*Rime*» (152)

io intendo scultura, quella che si fa per forza di levare:
quella che si fa per via di porre, è simile alla pittura

Michelangelo, Lettera a messer Benedetto Varchi

Me encantan estas dos citas de Miguel Ángel, porque dan la idea de que las cosas son lo que son, no sólo por lo que se añade (como se hace en la pintura), sino también por lo que se quita (como se hace en la escultura). El David, su David, estaba ya en el bloque de mármol cuando él llegó a esculpirlo, Miguel Ángel se limitó a quitar la piedra sobrante para que creciera la figura, «*una viva figura, che là più cresce u' più la pietra scema*». Una idea aplicable, igual que a todas las demás cosas de la vida, a esta tesis, que se pudo hacer gracias a la cantidad de cosas que se descartaron.

En esa línea, quiero agradecer a mi director de tesis, Jesús Vázquez. Igual que Miguel Ángel quitaba con su cincel toneladas de mármol, Jesús eliminó, con su visión de conjunto, toneladas de tiempo y energías dándole la vuelta a cuestiones que a la larga son irrelevantes, mostrándome cuáles eran las cosas más importantes que había que resolver cada vez que nos enfrentábamos a problemas nuevos en esta aventura. A su vez, le agradezco todos estos años en los que confió en mí para llevar a cabo el trabajo que aquí se presenta.

Agradezco a Pedro que me quitara la idea de hacer una tesis en otro lugar, y que me animara a regresar a España para trabajar juntos en estos proyectos. Él conoce mejor que ninguno de quienes han pasado por el laboratorio lo que significa llegar a la proteómica desde la física, y lo complicado, a la vez que emocionante, que puede ser el sentirse a la vez un extranjero académico y un aventurero en territorios inexplorados. No lamento los dolores de cabeza, que fueron mutuos, pues es gracias a ellos que mereció la pena.

Agradezco a mi codirectora de tesis y compañera de faenas Elena todas las horas que me evitó perder gracias a su iniciativa y a su increíble capacidad de trabajo, que fueron críticos para empujar los artículos que presento en este trabajo. Igualmente agradezco a mi compañero de penurias Fernando, que erradicó el desaliento que acechaba cada vez que se complicaban las cosas; el triángulo de la biología de sistemas fue una noble causa común por la que luchamos a brazo partido (algunos de manera más metafórica que otros). A Iakes le agradezco, además de su compañerismo, que me quitara la idea de hacer SanXoT con bases de datos desde el primer momento, lo cual ayudó a simplificar la tarea lo suficiente como para hacerla viable a corto plazo. Este párrafo no puede terminar sin que les dé las gracias a los demás compañeros del despacho, sin quienes el viaje habría sido, como mínimo, más solitario: Navratan, आपने मदद की है मुछे मेरे बचपन को याद करने मैं, धन्यवाद दोस्त; Juan Antonio, *madre mía, madre mía*, gracias por tantos buenos consejos; Enrique, por cuestionar la primera ley de la termodinámica y disfrutar de mis metáforas; Spyros, por tener siempre una sonrisa y crear buen ambiente.

También quiero recordar aquí a los compañeros con quienes empecé el viaje y que ahora están trabajando en otros lugares, como Pablo, gracias a quien se eliminaron tantos *bugs* de QuiXoT. A Daniel, por explicar tan bien las cosas que él sabía cuando yo estaba empezando. A Salva, que aunque no coincidimos en el tiempo, coincidimos en el pRatio. Y a Margoth, por contarme tantas cosas interesantes de su Perú natal.

No puedo dejar de mencionar a mis colegas del laboratorio, tanto quienes me acompañaron desde los tiempos del CBM, como Inma, Estefanía, y Raquel, como a quienes conocí más tarde, como Emilio, Aleksandra, Marta, Celia, Diego, Jesús (gracias por desempolvar el cálculo diferencial), Ricardo, Rocío, Mariano, Michał, Luis (que, aunque estuvo apenas unas semanas, con sus ganas de aprender logramos sacar adelante al MaesePedro), y Lenka (que vio las primeras versiones de SanXoT allá por 2012). Cuando empecé a escribir la tesis también me hicieron mucha compañía Alessia (grazie per avermi fatto sentire un po' a casa), Ileana, Lara y Víctor. También mencionar a Antonello (che mi ha fatto sentire un po' più in compagnia come fisico al CNIC), Moreno, Valeria y Verónica. Más allá, en el mundo de la proteómica, le agradezco a Anabel su energía y experiencia, a Margarita la sinergia en la distancia, a Gorka sus interesantes aportaciones que siempre nos tienen pensando, y a Jesús Jorrín el discurso que dio en las Jornadas de Proteómica de Córdoba. También un agradecimiento a Barrett Lyon y The Opte Project por autorizar el uso de su obra «Visualization of the routing paths of the Internet» para la portada de este trabajo (que considero apropiada, dado que las redes de interacción entre proteínas podrían seguir las mismas leyes que la red mostrada en dicha obra).

A los amigos que no saben de qué trata esta tesis, pero que me vieron tan ocupado con ella, les agradezco su paciencia, sobre todo este último año, por no haber podido verles todo lo que me hubiera gustado: David (un hermano de facto), Mónica, Angela, Rita (obrigado pela empatia durante a redação desta tese), Luis (tenemos pendiente quedar telescopio en mano), Edurne, Pablo, Alberto (siempre con temas de ecología y ciencia interesantes), Oli (añoro aquellas conversaciones sobre viajes), Rocío (que también sabe lo que es completar un libro), Sergio (a quien veo menos ahora, pero que siempre está en momentos clave), Rosa (gracias por el apoyo en los momentos difíciles), Paula (con quien siempre coincido, primero en la UAM, luego en Dublín, y ahora, de manera un poco más etérea, en bioinformática), Tomás (esos agradables paseos en bici por las mañanas), y a mi hermana Taína (que, desde tiempos inmemoriales, tiene un doctorado honoris causa en aguantarme).

A Cosme, que es el único aquí que estoy seguro de que no leerá esto, entre otras cosas por no ser humano, le debo mucho por haberme ayudado a combatir las preocupaciones con kilómetros y kilómetros de canicross, y por años de compañía silenciosa.

A Leonardo, que tantas ganas tiene de ver esta tesis, muchas gracias por los momentos de alegría de cada día. Si Miguel Ángel quitaba bloques de mármol, él quita montañas de grisura a los días, llenándolos de color. Estoy seguro de que con la curiosidad que tiene, destacará haga lo que haga.

A Matteo, que todavía no sabe a dónde voy todas las mañanas, le debo muchas horas de felicidad. Siendo como es, el más sociable de la familia y siempre cargado de buena intención, sin lugar a dudas cargará de entusiasmo a las personas que le rodeen.

Y a Marga, que por esta tesis ha hecho tantos sacrificios como yo, le debo más que a nadie. De hecho, este trabajo representa casi las tres cuartas partes del tiempo que llevamos juntos. Muchas gracias por tu compañía todo este tiempo, y el que quede.

Abstract

The analysis of high-throughput proteomics data presents the challenge of extracting biological meaning from a wealth of protein identifications and quantifications. In the last decade, technology in this area has undergone a major transformation that required a continuous and enormous development of bioinformatic tools to establish the foundations of the algorithms to be used in the next years of proteomics research. In this work we present three papers that represent three milestones in this endeavour.

In the first publication, we present a deep analysis on the performance and influence of the peptide identification search algorithms upon the appearance of high-resolution and high-accuracy mass spectrometres. It is shown that, in many relevant cases, using smaller precursor ion mass tolerances to identify peptides leads to an increased number of incorrectly identified peptides greatly underestimated by the false discovery rate (FDR). Here we propose a change in the search algorithm, consisting of the use of wide mass windows followed by a post-scoring mass filtering.

The second publication is dedicated to the WSPP (initialism for Weighted Spectrum, Peptide, Protein) statistical model for the analysis of high-throughput quantitative proteomics experiments. The model can be used in a wide range of combinations of stable isotope labelling (SIL) techniques and mass spectrometres. Additionally, this algorithm provides a general statistical framework for these experiments, allowing the comparison of results across laboratories, thanks to its unique capacity to separate the different sources of variance, allowing the interpretation of the error at different levels.

In the third and final paper, we present an innovative method to perform systems biology analyses from the proteomics perspective, considering the degree of coordination of a proteome, and thanks to the statistical basis provided by the WSPP statistical model. This was possible after developing the Generic Integration Algorithm (GIA), which allowed integrating quantitative information from any lower level to any higher level (instead of limiting us to the traditional spectrum-peptide-protein workflow). All these models are implemented in SanXoT, a software package developed to allow the practical use of the mentioned models in quantitative proteomics.

These three steps in the research of high-throughput proteomics represented a dramatic change in the way proteomes were analysed in our laboratory, and opened countless possibilities for further development and enhancement of this research topic.

Resumen

El análisis de datos en proteómica de alto rendimiento lleva implícito el reto de extraer significado biológico a partir de un gran número de identificaciones y cuantificaciones. En la última década la tecnología en este campo ha sufrido una transformación sin precedentes; esto ha requerido un desarrollo colosal de las herramientas bioinformáticas para establecer los fundamentos de los algoritmos que se usarán en los próximos años de investigación en proteómica.

En la primera publicación, analizamos en profundidad el rendimiento e influencia de los algoritmos de identificación de péptidos tras la aparición de la espectrometría de masas de alta resolución. Se muestra que, en muchos casos, reducir la tolerancia de la masa del ion precursor para identificar péptidos nos lleva a un aumento en el número de péptidos identificados incorrectamente, subestimados en gran medida por la tasa de error FDR[1]. Proponemos aquí un cambio en el algoritmo de búsqueda, consistente en el uso de ventana ancha para la masa del precursor, seguida de un filtrado de dicha masa tras calcular la puntuación asignada a la identificación.

La segunda publicación trata del WSPP [2], un modelo estadístico que desarrollamos para el análisis de experimentos de proteómica cuantitativa de alto rendimiento. Se puede utilizar en numerosas combinaciones de métodos de marcaje isotópico estable (SIL) y espectrómetros de masa. Además, aporta un marco estadístico general para estos experimentos, permitiendo la comparación de resultados entre distintos laboratorios gracias a su capacidad única para separar las diferentes fuentes de varianza, así como la interpretación de resultados a distintos niveles.

En el tercer y último artículo presentamos un método innovador para el análisis de biología de sistemas en proteómica, aprovechando la base estadística del WSPP, y teniendo en cuenta el grado de coordinación del proteoma. Esto fue posible gracias al desarrollo del Algoritmo de Integración Genérico (GIA), que permitió que la información cuantitativa se pudiera integrar desde cualquier nivel inferior a cualquier nivel superior (en vez de limitarnos a la secuencia estándar espectro-péptido-proteína). Todos estos modelos están implementados en SanXoT, un paquete de *software* que pone en práctica los modelos mencionados para proteómica cuantitativa.

Estos tres pasos representaron un cambio drástico en los métodos empleados para analizar proteomas en nuestro laboratorio, y abren la puerta a infinidad de posibilidades para futuros desarrollos y mejoras en proteómica de alto rendimiento.

[1] Sigla en inglés para *False Discovery Rate*
[2] *Weighted Spectrum, Peptide and Protein*, en español *Espectro Ponderado, Péptido y Proteína*

Table of contents

Sections in English: green colour
Apartados en español: de color azul

Índice de apartados en español

Abbreviations

term	description
AngII	angiotensin-II
ANOVA	Analysis of Variances
CA	category-to-all (mostly used in subscripts)
CE	chicken embryo extract
CORUM	Comprehensive Resource of Mammalian protein complexes
DB	data base
DIA	data-independent acquisition
DNA	deoxyribonucleic acid
FBS	foetal bovine serum
FCS	functional class-scoring
FDR	false discovery rate
GIA	Generic Integration Algorithm
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HDAC6	histone deacetylase 6
HDL	high-density lipoprotein
HPLC	high-performance liquid chromatography
HS	human serum
HT	high-throughput
IL-2	interleukin-2
IPA	Ingenuity Pathway Analysis
iTRAQ	isobaric Tags for Relative and Absolute Quantitation
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	knock-out
LC-MS	liquid chromatography coupled to mass spectrometry
LF	label-free
LTQ	linear trap quadrupole
MAM	mitochondrial-associated endoplasmic reticulum-membrane
MCL	Markov Clustering Algorithm
MCODE	Molecular Complex Detection
mRNA	messenger ribonucleic acid
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MS^1	survey scan (or full scan)
MS^2	fragmentation spectrum
ORA	over-representation algorithm
PANTHER	Protein ANalysis THrough Evolutionary Relationships
PMF	peptide mass fingerprinting
ppm	parts-per-million

term	description
PRDX6	peroxiredoxin-6
PSM	peptide-spectrum match
PTM	post-translational modification
RNSC	Restricted Neighborhood Search Clustering
QA	protein-to-all (mostly used in subscripts)
QC	protein-to-category (mostly used in subscripts)
QP	quantitative proteomics
RNA	ribonucleic acid
RPMI	Roswell Park Memorial Institute medium (RPMI medium)
SB	systems biology
SBT	Systems Biology Triangle
SIL	stable isotope labelling
SILAC	Stable Isotope Labelling by Amino acids in Cell culture
SNP	single-nucleotide polymorphism
SPC	Super Paramagnetic Clustering
TMT	Tandem Mass Tags
VSMC	vascular smooth muscle cell
VSN	Variance-Stabilising Normalisation
WSPP	Weighted Spectrum, Peptide, Protein [statistical model]
WT	wildtype

Introduction

> We cannot define anything precisely. If we attempt to, we get into that paralysis of thought that comes to philosophers, who sit opposite each other, one saying to the other, "You don't know what you are talking about!". The second one says, "What do you mean by *know*? What do you mean by *talking*? What do you mean by *you*?"
>
> Richard Feynman, «*The Feynman Lectures on Physics*», volume I

1. Overview

1.1 Area covered by this work

The present work treats the area in the intersection of several research fields, especially proteomics, statistics and bioinformatics. All these fields have had a major boost in the last ten years: proteomics deals with the study of the proteome in a biological system, which means detecting, quantifying and analysing the proteins expressed by the cells of the system under study; statistics is the branch of mathematics that studies and characterises sets of data in order to infer conclusions about the nature of the system related to these data; and, finally, bioinformatics is the emergent, interdisciplinary field using, as a bridge, computational resources and algorithms to solve biological problems, such as those posed by proteomics, among others.

1.2 Three challenges for bioinformatics in proteomics research

The area covered by this work is a fertile terrain in need of new ideas (or novel applications of ideas already used in other fields) and, at the same time, an open space difficult to dominate. The main challenges are associated with the breathtaking pace driven by the new possibilities of biological discoveries, instruments, and computation. For example, if we accept that big data in genomics more than doubles every year (Marx, 2013), then, since this work started, the world has experienced at least an eight-

fold increase in genomic data; a data explosion difficult to conceive. But this pace is not only limited to the amount of data; for instance, accuracy of mass spectrometres improved at a pace the identification algorithms could not follow, and even though many more, better spectra were produced by new machines, and hence more peptides were identified (and subsequently, in more complex experiments, quantified), the number of *good* peptide identifications decayed. The method to calculate the FDR with new technology had to be modified, which led to the first of the three papers presented in this work. But after obtaining more, good quality identifications, we faced another limitation, which was to take advantage of the many quantifications at hand: we had more identifications, good identifications. But knowing a protein or a peptide is present is not enough, it is also important to know its relative abundance compared to other samples (such as control samples, or different patients) to understand what is happening in the proteome. However, increasing the number of identifications did not improve the information obtained from the quantifications of the proteins. For this reason, it was a priority altering the algorithm to perform the statistical analysis, so that the information at the protein level could be correctly interpreted. We further developed an existing statistical model to meet the research needs, and the result is presented in the second paper. Then, we faced a third limitation. We obtained many good peptide identifications, and many good protein quantifications. But obtaining many quantified proteins did not help understanding better what was happening in the proteome under study. Available systems biology algorithms, such as enrichment algorithms, did not take advantage of all the information provided by the workflow used. To improve the conclusions drawn from the lengthy tables of quantified proteins, we developed a definition of protein coordination, which is a central concept in the Systems Biology Triangle, and is presented in the third and final paper within this work. Using this algorithm, we were, for the first time, able to detect the proteins that were presenting a differential behaviour compared to other proteins in the same ontological context, and, at the same time, we were able to obtain category-level trustable information about proteomes, replacing lists of thousands of proteins by lists of few, but highly relevant, categories affected.

As in every research project, the work does not finish here. New challenges arise after solving the old ones. However, this work represents three different steps of improvement that have been critical for the advancement of the research in our laboratory during the last years.

2. Identification of proteins

2.1 Peptide-centric MS/MS-based identification of proteins

The most logical starting point in the broad field covered by this work is taking into account the current techniques to identify the proteins present in a biological sample. To be able to identify a protein, a number of protocols have been developed to extract, isolate, digest and analyse the result. Here we will focus on the computational aspect.

Several analytical techniques are used to identify the amino acid sequence that defines a specific protein; one possibility is sequencing them de novo, especially when there is no prior knowledge of the sequence, although for practical reasons the main approach to reveal the protein content of a sample is by comparing the spectra from a mass spectrometre to a prior database using one of the different algorithms available.

One of the first approaches consisted in simply measuring the mass of the ions detected after digesting the proteins into peptides with a protease (mostly trypsin), and then using a mass spectrometre to detect their masses. As peptides have a very specific sequence, their mass, which is the sum of the masses of their amino acids, is also very specific and in certain situations it can be a good starting point to identify a protein. This led to the peptide mass fingerprinting technique (PMF), consisting of digesting the protein, obtaining a list of masses of its peptides, and comparing these with a prediction of the peptides available from a large set of protein candidates which have been digested *in silico* (Henzel, 1993; James, 1993; Mann, 1993; Pappin, 1993; Yates, 1993). However, this technique has been superseded with the advent of high-throughput proteomics.

Peptide mass fingerprinting is subjected to several ambiguities, for example due to permutations within the sequence (which do not change the mass) or the mass of certain amino acids (which is the case of leucine vs isoleucine, with identical mass, or lysine vs glutamine, whose mass is very similar). This and other factors show that PMF relies heavily on the size, abundance and amount of isolated peptides, which is scarce to give enough information to provide an acceptable certainty about the identification of a protein; for example, if only two peptides are found from a certain protein, this would give only two pieces of information. These ambiguities are manageable when considering few proteins in each sample and few candidates while having a large amount of data to back them. But when dealing with thousands of proteins, some of them with few peptides, it becomes clearly insufficient. Hence, this technique was replaced by the fragmentation method (Mortz, 1996), in which peptides, instead of proteins, are identified. It consists of using fragmentation spectra (also called MS^2, or

MS/MS spectra) to identify a single peptide, which are obtained by using two steps in the identification process: first, isolating the whole-peptide ion (called precursor ion), and second—using different techniques, but mainly collisions with particles at high kinetic energy in the medium—breaking that ion into fragments, many of which keep at least one charge that allows their mass to be measured by the spectrometre. The method is much more specific than PMF, as, instead of having a single value for each peptide (its mass), fragmentation spectra supply a set of values for it that can be compared with predictions performed by an algorithm making use of large databases.

In most cases, precursor ions are broken in two at their backbone, especially at their peptide bond (but also before or after it). To identify the resulting fragments, the Roepstorff-Fohlman notation is commonly used (Roepstorff and Fohlman, 1984) (Figure 1); in this notation, fragments are named so that the ions keeping the N-terminus are labelled *b* (if the peptide broke at the peptide bond), *a* (if it broke at the previous bond in the backbone) or *c* (if it broke just after the peptide bond); while fragments keeping the C-terminus are labelled *y* (when broken at the peptide bond), *x* (broken before it) or *z* (broken after it). The most common fragments observed are *b* and *y*.

The algorithm used to match each experimental spectrum to the most plausible peptide in the database is of key importance in high-throughput proteomics, as it is the first step of any analysis. The comparison is usually reduced to a single score. Scores

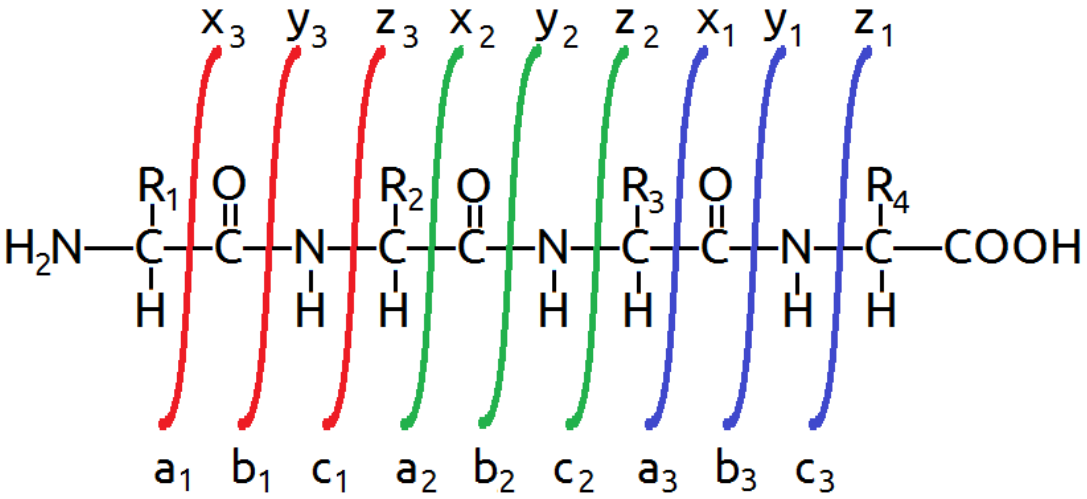


Figure 1: Schema of the Roepstorff-Fohlman notation for peptide fragments. Fragment ions receive a different name, according to the side of their side of the peptide (either N- or C-terminal), the number of amino acids kept from the original peptide, and the bond broken during fragmentation. Ions *a*, *b* and *c* are the fragment ions on the N-terminal side, while *x*, *y*, and *z* ions are the fragment ions on the C-terminal side. Subscripts show the number of amino acids that are kept.

can be either 1) parametric (or independent), when the algorithm takes into account only the experimental spectrum, comparing it to an *in silico* spectrum from the database (so that spectra and peptide candidates are considered individually: the score resulting from matching an experimental spectrum to the predicted spectrum of a peptide candidate is independent of other spectra or peptide candidates), or 2) probabilistic (or DB-dependent), when the a whole set of spectra (either experimental, or theoretical, or both) is used to obtain a global distribution and calculate the probability of a PSM of being a bona fide identification. For example SEQUEST's XCorr and Mascot's ions score fall in the first group (Yates, 1995a; Yates, 1995b; Cottrell and London, 1999), while the parametres from X!Tandem (Craig and Beavis, 2003, 2004), pRatio (Martínez-Bartolomé, 2008), and the other parametre from Mascot (the expectation value) are associated to the latter (Cottrell and London, 1999).

2.2 Concept of false discovery rate (FDR)

Statistical hypothesis testing is the branch of statistics dealing with measuring the certainty of scientific statements. Current mass analysers provide a huge amount of data, but a set of data is itself worthless if the conclusions we infer from it are wrong or incorrect, or, even worse, if we do not know if they are incorrect or not. There is a need to "feel" the reliability of the information we are handling.

The correctness of data shapes the knowledge we can obtain from an experiment, and it is a central point in the problem of induction. Philosophers like David Hume or Karl Popper have discussed this problem far beyond the practical needs of a bioinformatician. In short, we can remark that Hume stated that inductive methods infer "instances of which we have had no experience resemble those of which we have had experience" (Hume, 1738), while, on the other hand, Popper claimed that induction has no function in a logic of science, so the truth of a theory had to be established by empirical evidence (Popper, 1963). This is important, as in the end led to Fisher's definition of null hypothesis in the context of statistical hypothesis testing: one can only prove or disprove a null hypothesis, there is nothing such as an absolute certainty; results can only be statistically significant (stand out), or not, compared to the null hypothesis (Fisher, 1935).

For the bioinformatician working on peptide identification, the null hypothesis considers the identifications provided by an algorithm as wrong, as if the peptides had been assigned randomly. As a starting point, we need a score to quantify the goodness of the identification. Every algorithm has its own score (such as SEQUEST's XCorr), but a score itself has no information, it is just a number which has to be compared to a distribution. The position of that number within the distribution will tell us if it is in the

extreme or in the bulk of the distribution. A common way to observe the behaviour of data when identifications are wrong is performing a search with a decoy database made of random peptide sequences, and comparing the distribution of scores there with the scores obtained in the target search. By direct extrapolation from the distribution of decoy search scores, we can obtain the *p*-value, which is the probability of getting the score (or a more extreme score) we are obtaining in the target search just by chance. This concept was introduced first by Karl Pearson in 1900 (Pearson, 1900), and was popularised by Ronald Fisher in 1935 in his famous "lady tasting tea" experiment (Fisher, 1935). In peptide identification, if an assignation has an exceptionally good score, then we say the result is statistically significant, and we can reject the null hypothesis (which asserts the identification is false, brought just by chance). This means that we either have a bad identification with a very improbable high score, or, alternatively, we have an identification that does not fulfil the null hypothesis (i.e., the assignation has not been produced by chance, but because the peptide is really producing the spectrum observed, and therefore the identification is correct). The significance level at which a null hypothesis should be rejected, or not, depends on the experiment and the needs of the researcher. Obviously, there is a compromise between the degree of certainty and the amount of data that fulfils the requirements.

Unfortunately, the *p*-value is not the best choice for experiments having thousands of assignations, which are commonplace in high-throughput proteomics, as we have to face the multiple comparisons problem: while the *p*-value can be a good statistical tool for individual identifications, it becomes very impractical when we take into account large data sets, as we can easily obtain many false positives just by chance, in many cases even being more false positives than true positives. This problem became a serious concern in the 1980s and early 1990s, when researchers, mainly working in genomics, started to deal with very large amounts of data. To solve the problem, Yoav Benjamini and Yosi Hochberg presented in 1995 the concept of false discovery rate, or FDR, which provides the proportion of false hits present in a set of data (Benjamini and Hochberg, 1995). The most simple and commonly used version consists in simply getting, for a specified score threshold, the ratio N_{decoy}/N_{target} of hits whose score is better than the specified threshold.

The FDR is currently the most used parametre to measure the trustworthiness of a data set. As its use is very widespread, it has become a good mathematical tool to compare results between research groups worldwide. Nowadays, most peer-reviewed journals request FDR rates of 1% in the experiments performed, although sometimes the researcher might need a looser or stricter significance, in some cases for practical

reasons (for example, higher safety requires stricter significance levels, while lower costs are associated to looser standards), but also depending on what conclusions are to be demonstrated, sticking to the principle stated by Pierre-Simon Laplace, rephrased and popularised by Carl Sagan as "extraordinary claims require extraordinary evidence." (Sagan, 1980)

Different approaches have been published about the best way to calculate the FDR (Käll, 2007; Jones, 2009; Jeong, 2012). Starting by the way to generate a decoy database (which can be generated by reversing whole protein sequences, or reversing only peptide sequences maintaining arginines and lysines, or just shuffling the amino acids), the way to perform the search (in some cases, two searches are performed, one using the decoy and another for the target database, while in other cases one single search is done, by concatenating both databases and performing one single search), or the formula used to calculate the FDR—in which case some use the abovementioned N_{decoy}/N_{target}, but other researchers are more prone to use $2\,N_{decoy}/(N_{target}+N_{decoy})$, or even more refined methods (Navarro and Vazquez, 2009). Additionally, other parametres must be considered, such as the size of the database used (which is greatly affected, for example, by search parametres such as the precursor tolerance). All these versions and parametres would not be of great importance if their output was similar in all cases. However, the resulting list of identified peptides can vary greatly (Jeong, 2012), and, more importantly, the correctness of the output depends on the conditions and properties of the experiment: are there many scans or only a few? Is the resolution of the instrument high or low? Practical questions arise as well: is the algorithm slow or fast? Does it need many computational resources? Hence, deciding between algorithms and parametres is not a trivial task.

2.3 False discovery rate estimation problems associated with the use of narrow mass precursor windows

The improvements in liquid chromatography coupled to mass spectrometry (LC-MS) made in recent years have boosted the use of this technique for the characterisation and quantification of protein components in biological systems. The identification of thousands of proteins is today common practice, even in single HPLC runs. This enhanced degree of proteome coverage is mainly due to improved scanning speed and detection sensitivity, but also reflects increases in resolution and mass accuracy. For instance, resolutions of up to 480,000 have been described for approaches using modern Orbitrap™ mass analysers (Hebert, 2013). Better mass accuracy allows database searches to be performed at narrower precursor mass

tolerances, lowering the number of potential sequence candidates. Increasing mass accuracy is therefore widely assumed to diminish the chance of false assignations, thereby increasing the specificity of peptide identification. The notion that the quality of peptide identification is fully controlled in any condition by estimating the false discovery rate (FDR) using the decoy-target approach has encouraged scientists to perform database searches at increasingly narrower precursor mass tolerances. However, as it has been anticipated in section 1.2, little attention has been paid to the fact that by narrowing the mass tolerance the number of sequence candidates can diminish markedly, limiting the amount of data available for identification statistics. In one of the pioneering studies using the target-decoy strategy (Elias and Gygi, 2007), it was already recognised that estimates of the number of false-positive peptide-spectrum matches (PSMs) are less accurate when high precursor-mass-accuracy searches are performed with relatively few candidate sequences. Subsequent studies have warned of the possible unreliability of scores that implement certain elements of probabilistic modelling in the case of highly constrained searches (Nesvizhskii, 2010). More recently, questions were raised about the accuracy of peptide identification when using very narrow parent ion mass tolerance (Cooper, 2011; Cottrell and Creasy, 2011; Cooper, 2012; Chalkley, 2013). We first noted this problem when we began to use instruments with higher mass accuracies than the linear ion traps we used previously, especially after solving limitations of the algorithms used for quantifications by developing a statistical model (Jorge, 2009; Navarro, 2014) that was able to robustly detect quantitative peptide outliers (peptides whose quantification deviates from that of other peptides from the same protein). We observed that narrowing the mass precursor tolerance for peptide identification increased the number of identified peptides, but also increased the number of quantitative peptide outliers, suggesting a concomitant decrease in the reliability of peptide identification (unpublished results).

Interestingly, the use of narrow precursor mass windows does not necessarily improve peptide identification. A number of studies have shown that identification performance can be increased by using a database search with a wide mass tolerance followed by a posterior filtering or reanalysis that takes into account the mass deviation between the experimental and the theoretical mass (Beausoleil, 2006; Brosch, 2008; Ding, 2008; Hsieh, 2010; Nesvizhskii, 2010). This approach also avoids lowering the number of sequence candidates and hence potential problems derived from the use of small search spaces. Unfortunately, this method increases search times and is not used in most commercial search packages. Thus despite the concerns expressed in the literature, the high mass-accuracy approach for database peptide searching is now widespread (Kim, 2014; Wilhelm, 2014). The first paper presented in this work will present a solution to these problems.

3. MS/MS-based quantification of proteins

3.1 Relevance of protein quantification

Biological systems are dynamic and contain many molecular species (including DNA, RNA, proteins, carbohydrates and lipids) whose interactions result in complex series of physicochemical, spatial and temporal changes. Since proteins participate in all cellular processes, knowing how protein amounts change over time, along with their activity, provides important information about the state of the system. It is from such knowledge that the molecular mechanisms of disease and new pharmacological targets and biomarkers eventually emerge. Recent advances in mass spectrometry (MS) based proteomics allow the identification and relative quantification of thousands of proteins in a single study. Despite these advances, the reproducibility of MS-based proteomics has been called into question (Aebersold, 2009). It is an accepted fact that when the technology is properly applied, it is highly reproducible (Nilsson, 2010); therefore, progress in the field will depend on a correct understanding of these techniques and their limitations (Domon and Aebersold, 2010). The development of suitable statistical models is a critical step toward achieving this goal.

3.2 Label-free approaches

MS-based quantitative proteomics may be performed by direct quantification of precursor or fragment ion peptide intensity in each of the samples (label-free approaches), or by using stable isotope labelling (SIL) techniques. In the most common setup, label-free quantification involves analysing several technical replicates of the same sample or samples from several subjects (biological replicates) belonging to two or more different conditions. Due to the multiplicative nature of the different factors (fixed effects) and error sources (random effects) involved, quantitative data are usually subjected to a logarithmic transformation that allows treating these effects as additive, providing a natural way for modelling the replicate structure of the data within the Analysis of Variance (ANOVA) framework (Clough, 2009). A common feature of these ANOVA models is that the replicated structure are used to estimate the fixed effects and the variances associated to random errors, which are assumed to be normally distributed (Daly, 2008; Polpitiya, 2008; Clough, 2009; Karpievitch, 2009; Oberg and Vitek, 2009; Chang, 2012). Although the analysis may be performed at the peptide feature level (one test per feature) (Daly, 2008; Clough, 2009), the quantitative results from different peptide features belonging to the same protein are usually integrated to make the analysis at the protein level (one test per protein). As it has been explained in section 2.2, and taking into account that the analysis is repeated for very large

numbers of proteins (multiple hypothesis testing), statistically significant protein abundance changes in the different conditions are then detected by adjusting the *p*-value threshold to control for the false discovery rate (FDR) (Benjamini and Hochberg, 1995). In these models all peptide features are considered to contribute equally to protein abundance, which may be estimated from the plain average of feature log-corrected intensities (Clough, 2009), or from features corrected by fixed effects (Daly, 2008; Clough, 2009; Karpievitch, 2009) or scaled up to the same level (Polpitiya, 2008) before making the protein average. In these approaches, random errors are assumed to derive from only one source, so that the variance is a measure of the technical variability, but in some cases the biological variability is also taken into account by decomposing the total variance into the biological and the technical components (Daly, 2008; Clough, 2009).

3.3 Stable isotope labelling approaches

Stable isotope labelling (SIL) techniques—including stable isotope labelling with amino acids in cell culture (SILAC) (Ong, 2002; Ong and Mann, 2006), isobaric tagging for relative and absolute quantification (iTRAQ) (Ross, 2004), tandem mass tags (TMT) (Thompson, 2003), and enzymatic $^{16}O/^{18}O$ labelling (Mirgorodskaya, 2000; Bonzon-Kulichenko, 2011b)—currently offer the most accurate means of performing comparative quantitative proteomics studies (Cox and Mann, 2007). Although the existence of separate technical, experimental (Daly, 2008; Clough, 2009) and biological variations in iTRAQ has been analysed (Gan, 2007), no statistical models were derived from these studies. The simplest model to analyse iTRAQ data is to calculate the protein value as an average of peptide-ratios and compare each of the protein values across several replicates (one test per protein) using an appropriate statistical test such as Student's t-test. The protein average may be calculated using the mean (Unwin, 2005), the median (Boehm, 2007; Rodriguez-Suarez, 2010) or an average calculated minimising the square root distance from the peptide readings from the log-transformed peptide ratios (Boehm, 2007). This kind of testing at the protein level has been extended using an ANOVA model, similar to those proposed to treat label-free data, with additive peptide effects and only one random effect that combines the biological error and the measurement noise (Hill, 2008; Oberg, 2008; Oberg and Mahoney, 2012; Herbrich, 2013), which has been applied to the analysis of a case study involving four treatment groups with several replicates each (Oberg and Vitek, 2009). All these approaches have in common that all the peptide readings originating from the same protein are equally considered, under the implicit assumption that they have the same variance. This assumption is based on original analyses showing that non-corrected ratios of peptides measured by iTRAQ follow approximately a log-normal distribution

(Boehm, 2007). However, other analyses have demonstrated that not all the peptides are quantified with the same accuracy, existing a clear dependence of variance with ion intensity (Shadforth, 2005; Lin, 2006; Bantscheff, 2008; Karp, 2010; Zhang, 2010), which may produce deviations from normality. Therefore, it has been proposed using intensity-weighted peptide averages of log-ratios (Lin, 2006; Mahoney, 2011) to calculate protein averages. Other approaches try to model or control the behaviour of variance by a two-parametre modelling of the dependence of peptide variance with intensity (Zhang, 2010), or using a variance-stabilising normalisation (VSN) transformation (Karp, 2010; Arntzen, 2011)—similar to those employed in microarray approaches (Huber, 2002)—, calculating the protein values as the median (Arntzen, 2011) or the trimmed average (Karp, 2010) of transformed peptide values. It has been proved that the transformed ratios at the spectrum level have an apparently normal distribution, from which it is possible to detect statistically significant regulation of specific peptides (such as tyrosine-phosphorylated peptides) from the global distribution of transformed peptide ratios (Zhang, 2010). To test significance of the changes at the protein level, most methods assume that the set of protein values follow a normal distribution and use a 0.05 probability threshold (Lin, 2006; Arntzen, 2011). This test may be performed by direct fitting to a normal distribution (Lin, 2006), or using estimates of the standard deviation (Arntzen, 2011), while other approaches adjust the significance threshold so that 95% of experimental variation is encompassed (Karp, 2010). Finally, some authors correct the significance value for multiple hypothesis testing (Arntzen, 2011). Concerning SILAC data, the majority of studies use MaxQuant algorithm (Cox and Mann, 2008) to analyse quantitative results. In MaxQuant protein ratios are calculated as the median of all SILAC peptide ratios, and the proteins are then grouped into bins according to the sum of peptide intensities; in each bin protein log-ratios are then assumed to be normally distributed and the standard deviation is calculated using a robust estimate, from which a statistical significance is assigned to each protein. This procedure takes into account, empirically, the observed fact that highly abundance protein values have a lower variability than low abundance ones (Cox and Mann, 2008). Finally, we have proposed a statistical model to analyse quantitative data obtained by ^{18}O labelling, which decomposes the total technical variance into the spectrum, peptide and protein variance components (Jorge, 2009). This approach models the heterogeneous variance at the spectrum level and integrates log-ratios to the protein level using weighted averages according to error propagation theory. The validity of the model was demonstrated by showing that the distribution of protein values follows a normal distribution and by the very low percentage of outliers found at the spectrum, peptide and protein levels (Jorge, 2009).

3.4 The need for a universal statistical model for SIL approaches

All these studies show that, in spite of the efforts made in the field, a comprehensive statistical theory for the general analysis of quantitative data by SIL technique has not still been developed. Existing models are highly specific to each SIL method and mass spectrometre, making them unsuitable for examining data from different laboratories, judging experimental quality on the basis of unified criteria, handling, comparing and integrating multiple measurements, or interpreting the complete set of experimental results from different SIL approaches as a whole. Moreover, most models and statistical significance tests are based on normality assumptions that have not been tested, despite the fact that heterogeneity of variance has been documented in all SIL methods (Cox and Mann, 2008; Jorge, 2009; Karp, 2010). These techniques are based on peptide-centric measurements, and the lack of general models leads to the subjective choice of a method for combining multiple peptide readings to estimate protein ratios (Karp, 2010). This problem is further aggravated by the undersampling that characterises SIL-based MS analyses (Nilsson, 2010): the number of peptides that quantify a protein is variable and cannot be controlled between experiments—a non-trivial fact that complicates mathematical modelling. Finally, the lack of statistical standards hinders the development of systems biology tools to interpret quantitative proteomics data from more functionally-integrated perspectives.

Several of the problems posed here were solved in part in this work, as a continuation of the universal statistical model developed previously (Pedro Navarro, PhD Thesis). The results are presented in the second article of the *Results and Material & Methods* chapter and discussed in section 2 of the *Discussion* chapter.

4. Systems biology in proteomics and the concept of protein coordination

4.1 What do we intend by *systems biology*?

A system-level interpretation of biology has attracted interest since mid of the 20th century (Wiener, 1948). However, it can be considered an emerging discipline of the last fifteen years (Kitano, 2001), boosted especially after the completion of the Human Genome Project in 2003 (Ideker, 2001). And, as in every emerging discipline, especially interdisciplinary disciplines as this one (Aderem, 2005), there is no sharp definition about what is systems biology (Kirschner, 2005). However, in the last ten

years there is a growing consensus about its place within the bioscience research. For example, it can be said that both systems biology and molecular biology study the same problems (biology, biomedicine, biotechnology, etc) but from different approaches: molecular biology uses a bottom-up approach, by paying attention to the microscopic molecular details in the first place, and then considering larger structures (such as protein complexes, cellular components, membranes, organelles, cells, metabolic pathways, signalling pathways, tissues, organs, and eventually organisms and ecosystems); from this point of view, it is reasonable considering that systems biology treats the problem from the other side, using a top-down approach: starting with the macroscopic observed complexity of organisms (such as the performance of a specific signalling pathway, or the reaction of a tissue to a drug) the researcher uses mathematical models (like statistics, or graph theory) to individualise the different properties and causes that are behind, going down in the complexity scale, up to the molecular level (Kitano, 2002; Westerhoff and Palsson, 2004). These two approaches for the same problem can be observed also in other areas of science, such as, for example, thermodynamics vs statistical mechanics: thermodynamics uses a top-down approach deduced by the empirical observation of the macroscopic world, while statistical mechanics uses a bottom-up procedure, starting with the understanding of the different microstates of a system and how they explain the observed macrostate. From this point of view, statistical mechanics is to molecular biology as thermodynamics is to systems biology.

Another way to look at this is by considering that systems biology is—like data-mining approaches or machine learning methodologies—discovery-driven, instead of hypothesis-driven: in order to understand an extremely complex biological system, we perturbate it (biologically, genetically or chemically), we observe the response, and finally we create a mathematical model describing the system under study (Ideker, 2001). This means studying all the interacting elements at once, decoding a global response even when the details might remain unknown.

4.2 Systems biology: a multidisciplinary science

For its own nature, and the problems in question, systems biology is a multidisciplinary science (Aderem, 2005). Its common frame is biology, but the colossal amount of data requires several areas of mathematics and computer science to extract the relevant, meaningful information (Kitano, 2002; Stelling, 2004), and, as molecular biology itself, it is intertwined with biochemistry. The applications require a combination of the points of view of engineering and biomedicine. And, more specifically, systems biology is closely related to other multidisciplinary, emergent fields, not only related to

biological topics like genomics (Ge, 2003), proteomics (Weston and Hood, 2004; Cox and Mann, 2011), or metabolomics (Weckwerth, 2003; Kell, 2004; Tomita and Kami, 2012; Kanehisa, 2016), but also related to mathematics and computer science topics like graph theory (Mason and Verwoerd, 2007), network theory (Barabasi and Oltvai, 2004; Hood, 2004) or data mining (Ananiadou, 2006). The bridge between both sides, the biological and the mathematical, is one of the main areas of bioinformatics, a field that has been called “the computing response to the molecular revolution in biology” which “has reshaped the life sciences” (Finkelstein, 2004).

4.3 Systems biology models currently used in proteomics

Quantitative transcriptomics data are usually interpreted in relation to biological knowledge stored in databases, using a procedure known as gene set analysis or pathway analysis (Khatri, 2012). These knowledge databases may contain ontological information about genes, like Gene Ontology (GO) (Ashburner, 2000) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), or provide information about gene/protein interactions and how and where these occur, like Reactome (Joshi-Tope, 2005) or STRING (Szklarczyk, 2014). Pathway analysis algorithms that use ontological information are classified into two major subtypes: over-representation algorithms (ORA) and functional class scoring (FCS) (Khatri, 2012). ORA, also known as enrichment analysis, statistically evaluates whether the subset of genes showing significant expression changes relative to a given threshold is enriched in a given category (ontology) (Khatri and Draghici, 2005; Huang da, 2009). This approach is widely used to analyse quantitative proteomics data, but algorithms of this kind only consider significantly changing proteins and therefore ignore most of the acquired quantitative information; moreover, they do not consider protein abundance or fold-change information. In FCS methods, the quantitative values of all genes from a category are integrated to produce a category-level value (Khatri, 2012), which is statistically analysed to determine significant category changes (Mootha, 2003; Barry, 2005; Subramanian, 2005; Jiang and Gentleman, 2007). Although these threshold-free methods have also been used in proteomics and take account of all the information obtained, they were originally designed to treat transcriptomics data and therefore do not take optimal account of specific characteristics of protein quantification by mass spectrometry. Particularly, they do not consider the large dynamic range of protein concentrations typical of biological systems, which makes MS-based quantification of proteins present in low amounts very challenging and, in general, less reliable. This problem is aggravated by undersampling, whereby the number of peptides used to quantify a protein is variable and cannot be controlled (Navarro, 2014). Furthermore,

current FCS methods are not designed to analyse the presence and extent of coordination in protein behaviour.

4.4 The coordinated behaviour of proteins

Cellular processes are executed by proteins working together in complexes or functional pathways. The coordinated behaviour of proteins and their relations while performing their functions has attracted attention in recent times. This concept has been studied in both genes and proteins from different experimental and conceptual perspectives. High throughput mRNA analysis has revealed that genes that are functionally linked or are associated with the same metabolic pathway are often co-expressed (Ihmels, 2002; Ihmels, 2004; Wei, 2006), and that multi-protein complexes within the same functional class are regulated in the same direction (Sprinzak, 2009). By combining gene expression information with data about protein interactions, growth phenotype, and transcription factor binding, it was possible to describe groups of genes showing correlated behaviour (Tanay, 2004). Similarly, fluorescence techniques have revealed that transcript levels of temporally induced genes are highly correlated in individual yeast cells (Gandhi, 2011). At the protein level, high-throughput single-cell flow cytometry has been used to study biological noise (Newman, 2006), demonstrating that proteins which are subunits of the same complex tend to attain similar levels in the cell, that fluctuations in protein levels tend to be smaller within large complexes (Carmi, 2009), and that cell-to-cell variability in protein expression is similar for proteins sharing a similar function (Bar-Even, 2006). Using a yeast fusion library for immunodetection and measurement of absolute expression levels (Ghaemmaghami, 2003), it has been shown that interacting proteins are uniformly expressed (Carmi, 2006).

Recent works in MS-based proteomics have revealed that proteins with similar functions typically have similar expression levels (Marguerat, 2012). Similarly, a tendency of functionally related proteins to be co-ordinately regulated was demonstrated by correlation analysis of protein abundance after density fractionation (Foster, 2006), over a time course (Hansson, 2012), or in a large set of diverse conditions (Wu, 2014). In spite of these efforts, the mechanisms by which cells coordinate the levels of functionally related proteins are still poorly understood. Furthermore, existing methods to analyse quantitative data obtained from conventional proteomics studies (for instance when only two different conditions are compared) are not designed to detect coordination and least of all to analyse its extent.

Objectives

The rapid development of high-throughput proteomics, especially quantitative proteomics, has rendered many of the statistical models and computational tools obsolete. The impact has been catalysed by the general increase in the volume of data in the last years, the appearance on stage of high-accuracy mass spectrometry, and the increased prominence of systems biology in life sciences. These circumstances have left an ample void in the computational approaches used to date, and the main, general goal of this work has consisted in taking the bioinformatic resources and proteomic concepts to the next stage. Hence, the effort has been concentrated on the four specific areas that follow:

1. Generate novel algorithms to improve the calculation of the false discovery (FDR) rate in high-throughput proteomics identification experiments, paying particular attention to the artefacts exposed by the advent of high-accuracy mass spectrometry in the identification of peptides.

2. Adapt and improve the current statistical model used in our laboratory and its associated algorithm to allow its dynamical use in quantitative proteomics research, mainly by creating a general algorithm functioning for a wide range of workflows.

3. Establish concepts and algorithms that allow the analysis of the systems biology of pairwise quantitative proteomics experiments, paying special attention to the coordinated behaviour of proteins.

4. Develop the computational resources to allow the practical use of the concepts established in this work to facilitate the analysis of high-throughput proteomics experiments by third-parties.

Objetivos

El rápido desarrollo de la proteómica de alto rendimiento, y de la proteómica cuantitativa en particular, ha dejado obsoletos numerosos modelos estadísticos y herramientas computacionales. El aumento generalizado en el volumen de datos utilizados en los últimos años, así como la aparición de la proteómica de alta resolución y la importancia en aumento de la biología de sistemas en las ciencias de la vida, no han hecho sino catalizar esta situación. Estas circunstancias han dejado un vacío en las estrategias computacionales en uso hasta la fecha, de modo que el objetivo principal de este trabajo ha consistido en llevar los recursos en bioinformática y proteómica a la siguiente fase. Es por ello que el esfuerzo se ha concentrado en las siguientes cuatro áreas específicas:

1. Generar nuevos algoritmos para la mejora del cálculo de la tasa de error (FDR) en experimentos de identificación de proteómica de alto rendimiento, prestando especial atención a las deficiencias puestas de manifiesto tras la aparición de la espectrometría de alta precisión en identificación de péptidos.

2. Adaptar y mejorar el modelo estadístico en uso en el laboratorio, así como el algoritmo asociado, para permitir su utilización dinámica en la investigación en proteómica cuantitativa, haciendo especial énfasis en la creación de un algoritmo general aplicable a un amplio abanico de flujos de trabajo.

3. Establecer los conceptos y algoritmos que permitan el análisis de biología de sistemas en experimentos binarios de proteómica cuantitativa, prestando especial atención al comportamiento coordinado de las proteínas.

4. Desarrollar los recursos computacionales que pongan en práctica los conceptos establecidos en este trabajo y faciliten el análisis de experimentos de proteómica de alto rendimiento por parte de terceras personas.

Results and Material & Methods

When you're thinking about something that you don't understand, you have a terrible, uncomfortable feeling called confusion. It's a very difficult and unhappy business. And so most of the time you're rather unhappy, actually, with this confusion. You can't penetrate this thing. […] I get this feeling all the time: that I'm an ape trying to put two sticks together [to reach the banana]. So I always feel stupid. Once in a while, though, the sticks go together on me and I reach the banana.

Richard Feynman, 1963

1. First article

1.1 (English) Revisiting peptide identification by high-accuracy mass spectrometry: problems associated with the use of narrow mass precursor windows

As stated in the introduction, applying the narrowest tolerance window available for precursor ions to search MS/MS spectra through protein databases used to be the best option for low and medium resolution instruments. However, the advent of high-resolution mass spectrometry (with resolutions in the range of 30,000 to 500,000 or more) revealed a limitation in the minimum tolerance of precursor ions that can be used without decreasing the quality of peptide identifications by MS/MS.

In this paper we present a set of simple and generalisable exercises that pinpoint several problems related to reducing the number of sequence candidates. We differentiate between “independent” scores, which only depend on the peptide-spectrum match (PSM) that is evaluated, and “DB-dependent” scores, i.e. scores that take into account information from additional PSM candidates. Our results, obtained with two of the currently most used searching engines, SEQUEST and Mascot, suggest that estimation of FDR can be surprisingly accurate using independent scores and is not a problem in itself, even at very low precursor mass tolerances. However, we also found compelling evidence that FDR estimates may suffer from serious inaccuracies and that a considerable proportion of peptide identifications can be wrongly assigned when DB-dependent scores are used together with very narrow precursor mass windows.

My collaboration as third author consisted in i) the development of the algorithm for the post-scoring mass filtering, ii) its implementation in the software used to validate SEQUEST results (pRatio) (Martinez-Bartolome, 2008) (which I developed since version 5.1.4, in August 2009), iii) the participation in the interpretation of the results and iv) and the participation in writing the paper.

1.2 (Español) Nuevas observaciones acerca de la identificación de péptidos por espectrometría de masas de alta precisión: problemas derivados del uso de tolerancias pequeñas en la masa de los iones precursores

Tal y como se indica en la introducción, cuando se trabaja con instrumentos de resolución baja o intermedia, la mejor opción para buscar un espectro de fragmentación en bases de datos de proteínas suele ser emplear la menor tolerancia disponible para la masa del ion precursor. No obstante, la aparición de la espectrometría de masas de alta resolución (con la que se alcanzan resoluciones entre 30.000 y 500.000 o más) ha puesto en evidencia una limitación en la tolerancia mínima que se puede utilizar sin detrimento de la calidad de las identificaciones de péptidos.

En este artículo presentamos varios casos simples y generalizables que ponen en evidencia diferentes problemas que surgen al reducir el número de secuencias candidatas. Diferenciamos aquí entre *puntuaciones independientes* (que sólo dependen de la asignación péptido-espectro en consideración) y *puntuaciones dependientes* (que incorporan información adicional de otras asignaciones péptido-espectro). Nuestros resultados, obtenidos con SEQUEST y Mascot (dos de los principales motores de búsqueda), sugieren que la tasa de error (FDR) se puede estimar con una precisión y exactitud sorprendentes al utilizar puntuaciones independientes, incluso con tolerancias de masa de precursor pequeñas, no llegando a ser la causa de problemas de identificación. Sin embargo, también hemos hallado pruebas convincentes de que al combinar puntuaciones dependientes con tolerancias pequeñas para estimar la FDR se pueden cometer errores graves, derivando en una gran proporción de asignaciones péptido-espectro equivocadas.

Mi colaboración como tercer autor consistió en i) el desarrollo del algoritmo de filtrado de masas tras el cálculo de la puntuación de identificación, ii) su implementación en el *software* dedicado a validar los resultados de SEQUEST (pRatio) (Martinez-Bartolome, 2008) (de cuyo desarrollo me ocupé a partir de la versión 5.1.4, en agosto de 2009), iii) la participación en la interpretación de los resultados y iv) la participación en la escritura del manuscrito.

Journal of proteome research

Article

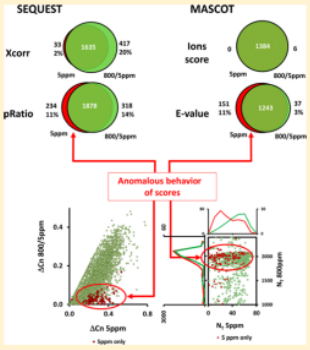
pubs.acs.org/jpr

Revisiting Peptide Identification by High-Accuracy Mass Spectrometry: Problems Associated with the Use of Narrow Mass Precursor Windows

Elena Bonzon-Kulichenko, Fernando Garcia-Marques, Marco Trevisan-Herraz, and Jesús Vázquez*

Laboratory of Cardiovascular Proteomics, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Melchor Fernández Almagro, 3, 28029 Madrid, Spain

ABSTRACT: Peptide identification is increasingly achieved through database searches in which mass precursor tolerance is set in the ppm range. This trend is driven by the high resolution and accuracy of modern mass spectrometers and the belief that the quality of peptide identification is fully controlled by estimating the false discovery rate (FDR) using the decoy-target approach. However, narrowing mass tolerance decreases the number of sequence candidates, and several authors have raised concerns that these search conditions can introduce inaccuracies. Here, we demonstrate that when scores that only depend on one sequence candidate are used, decoy-based estimates of the number of false positive identifications are accurate even with an average number of candidates of just 200, to the point that remarkably accurate FDR predictions can be made in completely different search conditions. However, when scores that are constructed taking information from additional sequence candidates are used together with low precursor mass tolerances, the proportion of peptides incorrectly identified may become significantly higher than the FDR estimated by the target-decoy approach. Our results suggest that with this kind of score the high mass accuracy of modern mass spectrometers should be exploited by using wide mass windows followed by postscoring mass filtering algorithms.



KEYWORDS: *False discovery rate, high-resolution mass spectrometry, peptide identification*

INTRODUCTION

The improvements in liquid chromatography coupled to mass spectrometry made in recent years have boosted the use of this technique for the characterization and quantification of protein components in biological systems. The identification of thousands of proteins is today common practice, even in single high-performance liquid chromatography (HPLC) runs. This enhanced degree of proteome coverage is mainly due to improved scanning speed and detection sensitivity, but also reflects increases in resolution and mass accuracy. For instance, resolutions of up to 480,000 have been described for approaches using modern Orbitraps.[1] Better mass accuracy allows database searches to be performed at narrower precursor mass tolerances, lowering the number of potential sequence candidates. Increasing mass accuracy is therefore widely assumed to diminish the chance of false assignments, thereby increasing the specificity of peptide identification. The validity of identification results is most commonly controlled by the false discovery rate (FDR), a generally accepted indicator that can be estimated using decoy databases (DB) in the form of composite[2] or separated target-decoy DBs.[3,4] The notion that the quality of peptide identification is fully controlled in any condition by estimating the false discovery rate (FDR) using the decoy-target approach has encouraged scientists to perform database searches at increasingly narrower precursor mass tolerances. However, little attention has been paid to the fact that by narrowing the mass tolerance the number of sequence candidates can diminish markedly, limiting the amount of data available for identification statistics. In one of the pioneering studies using the target-decoy strategy,[5] it was already recognized that estimates of the number of false-positive peptide−spectrum matches (PSMs) are less accurate when high precursor-mass-accuracy searches are performed with relatively few candidate sequences. Subsequent studies have warned of the possible unreliability of scores that implement certain elements of probabilistic modeling in the case of highly constrained searches.[6] More recently, questions were raised about the accuracy of peptide identification when using very narrow parent ion mass tolerance.[7−10] We first noted this problem when we began to use instruments with higher mass accuracies than the linear ion traps we used previously and to quantify results using a recently developed statistical model[11,12] that is able to robustly detect quantitative peptide outliers (peptides whose quantification deviates from that of other peptides from the same protein). We observed that narrowing the mass precursor tolerance for peptide identification increased the number of identified peptides and also increased the number of quantitative peptide outliers, suggesting a concomitant decrease in the reliability of peptide identification (unpublished results).

Interestingly, the use of narrow precursor mass windows does not necessarily improve peptide identification. A number of studies have shown that identification performance can be increased by using a database search with a wide mass tolerance

Received: July 11, 2014
Published: December 12, 2014

followed by a posterior filtering or reanalysis that takes into account the mass deviation between the experimental and theoretical masses.[6,13−16] This approach also avoids lowering the number of sequence candidates and hence potential problems derived from the use of small search spaces. Unfortunately, this method increases search times and is not used in most commercial search packages. Thus, despite the concerns expressed in the literature, the high mass-accuracy approach for database peptide searching is now widespread.[17,18]

In this article, we present a set of simple and generalizable exercises that pinpoint several problems related to reducing the number of sequence candidates. We differentiate between "independent" scores, which only depend on the PSM that is evaluated, and "DB-dependent" PSM scores, i.e., scores that take into account information from additional PSM candidates. Our results, obtained with two of the currently most used searching engines, SEQUEST and Mascot, suggest that estimation of FDR can be surprisingly accurate using independent scores and is not a problem in itself, even at very low precursor tolerances. However, we also found compelling evidence that FDR estimates may suffer from serious inaccuracies and that a considerable proportion of peptide identifications can be wrongly assigned when DB-dependent scores are used together with very narrow precursor mass windows.

EXPERIMENTAL SECTION

The test data set, containing 10,000 MS/MS spectra, was taken from an ongoing project and corresponded to the analysis of the ^{18}O-labeled mouse vascular smooth muscle cell nuclear proteome.[19] Peptides were RP-fractionated on a C-18 reversed phase (RP) Acclaim PepMap 100 nanoviper column (75 μm I.D., 50 cm, 3 μm, 100 Å) using a continuous acetonitrile gradient of 0−30% B for 4 h (B = 95% acetonitrile, 0.1% formic acid) at 200 nL/min on a nano-HPLC Easy-nLC 1000 interfaced with a Thermo Scientific LTQ-Orbitrap ELITE. Each MS cycle consisted of one MS scan in the orbital analyzer (390−1500 m/z, 240,000 fwhm at 400 m/z, profile mode, automatic gain control target of 1×10^6, maximum ion time of 50 ms, 2 microscans) followed by 20 data-dependent collision-induced dissociation MS/MS scans of the most intense peaks in the ion trap (10,000 ions target value, 1 microscan, 50 ms injection time, and 35% normalized collision energy). The dynamic exclusion was set at an exclusion window of ±10 ppm for 45 s. For peptide identification, MS/MS spectra were searched with Proteome Discoverer (version 1.3.0.339, Thermo Fisher Scientific) against different Uniprot databases (DBs) containing all mouse sequences (related DB, April 28, 2012), all *chordata* (larger related DB, March 20, 2013) sequences, or all mouse sequences doped with all *cyanobacteria* (larger nonrelated DB, April 28, 2012). The corresponding decoy DBs were created by inverting, in each protein of the target DB, the tryptic peptide sequences (except for the basic residues K and R, which were maintained in the same C-terminal position) and were searched separately. Search parameters were selected as follows: full or semitryptic digestion with up to 2 missed cleavage sites, 1,200 mmu fragment mass tolerance, carbamidomethyl cysteine as a fixed modification, and methionine oxidation and ^{18}O labeling on lysine and arginine residues as dynamic modifications. Precursor mass tolerance was 5, 10, 15, 25, 50, 100, 200, 400, 600, 800, 1200, 1400, 1600, 1800, 2000, and 3000 ppm. SEQUEST results were analyzed using the probability ratio method.[20] Database searches using Mascot were performed against the same decoy and target mouse DBs as for SEQUEST, using 5 and 800 ppm precursor mass windows and the same fixed and variable modifications. Postsearch result filtering by mass error was as described.[16] For each scan, if the mass deviation fell outside the ±5 ppm window, the corresponding XCorr or Ions score was rescored to 0, the pRatio was reassigned a value of 2, and the −log(E-value), a value of −10. The false discovery rate (FDR) was calculated as described.[21] The raw and identification data have been deposited to the PeptideAtlas with the data set URL identifier http://www.peptideatlas.org/PASS/PASS00529. The data from the reference mixture of Sigma UPS2 proteins (48 purified human proteins) analyzed on a Thermo Scientific LTQ-Orbitrap XL was downloaded from the Vanderbilt Mass Spectrometry Core Web site (http://massive.ucsd.edu/ProteoSAFe/status.jsp?task=f1aa4938fc2e4ee9b80d359b2df30f4b).

RESULTS AND DISCUSSION

FDR Can Be Estimated Accurately Using Decoy Databases Even at Narrow Precursor Mass Tolerances

We first studied whether it is possible to accurately estimate the number of false peptide−spectrum matches (PSMs) from decoy databases (DBs) when the number of candidates per spectrum is reduced significantly by narrowing as much as possible the precursor mass window according to the accuracy of the mass spectrometer that, in our case, was in the low ppm range. We used SEQUEST as a database search engine and the well-known XCorr score.[22] The XCorr score depends only on the MS/MS spectrum and the sequence with which it is compared, and is completely independent of the results obtained when comparing the same or other spectra with other sequence candidates. In this work we will refer to this kind of scores as "independent" scores, to differentiate them from "database-dependent" scores, which are the ones that take into account additional information from other sequence candidates present in the database against which the MS/MS spectrum is compared during the search. By using independent scores for this analysis we avoided the effect of DB-related confounding factors on the estimate of the number of false positives. We tested the validity of the basic principle of FDR calculation by the target-decoy approach: that all MS/MS spectra have the same chance of producing a random match with the same score in a target DB and in a decoy DB. We concentrated on the study of a representative collection of 10,000 MS/MS spectra generated from a mouse proteome; the number of fragment spectra in this collection is high enough to construct reliable score distributions, and at the same time, these numbers are easily obtained using modern mass spectrometers even in short runs. The spectra were searched at different precursor tolerances against two separate, nonrelated (NR) DBs that had no sequences in common with the mouse DB, so that all the MS/MS spectra were matched with false peptide sequences. The first NR DB, which we call NR-target, contained *cyanobacteria* sequences, whereas the second DB was the pseudoinverted version of the *cyanobacteria* peptides and was called NR-decoy. Precursor tolerances were 1, 5, 100, and 800 ppm, and produced an average of 227, 1,256, 19,296 and 62,038 candidate peptide sequences per spectrum, respectively. The score threshold for putative peptide identification was then moved from the lowest to the highest score, and at each score, the number of PSMs obtained against the NR-target DB was plotted against the number obtained against the NR-decoy DB, yielding a plot on which we could compare the results obtained in the two DBs for each of the tolerances. We observed that the number of "decoy" PSMs was identical to that of "target" PSMs at any score

dx.doi.org/10.1021/pr5007284 | J. Proteome Res. 2015, 14, 700−710

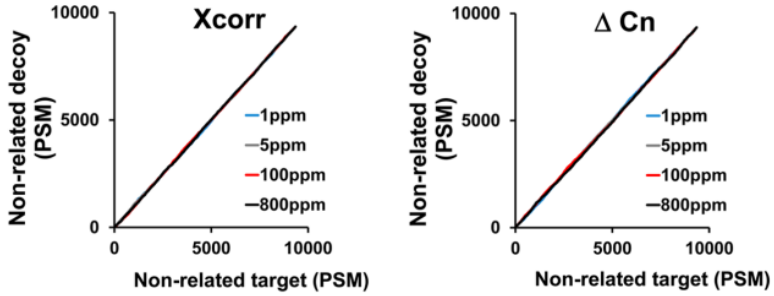


Figure 1. Accuracy of decoy estimations as a function of precursor mass tolerance. A collection of 10,000 MS/MS spectra from the mouse vascular smooth muscle cell nuclear proteome was searched against a nonrelated (cyanobacteria) DB at different precursor tolerances. The number of PSM obtained against the nonrelated target DB was compared with those obtained against the nonrelated decoy DB, using as filtering threshold score either XCorr (left) or ΔCn (right). The diagonal line represents the identity line.

threshold, independently of the precursor mass tolerance used (Figure 1, left). A similar result was obtained when ΔCn was used as the score threshold (Figure 1, right). Thus, in all cases the false PSMs are encountered with exactly equal probability in both DBs. This result seems to contradict previous findings reporting a noisier correlation between the number of target and decoy PSMs when the precursor mass tolerance was decreased.[5] However, the decoy DB in that study was constructed by reversing the sequence of the proteins, whereas here we reversed the peptide sequences in each protein, maintaining the C-terminal position of basic residues in each peptide so that the number of sequence candidates for each spectrum and the peptide mass distributions were exactly identical in the target and in the decoy DBs. Our results thus indicate that, when pure scores are used, the estimate of the number of false identifications using a decoy DB, and by extension the accuracy of the FDR estimate, is unaffected by the number of sequence candidates even at low precursor mass tolerances with as few as 200 candidates per spectrum.

FDR Behavior Is Predictable When Database Searches Are Performed in Different Conditions

It is well-known that the number of peptide sequence candidates used in a database search critically affects the number of false PSMs at a fixed peptide identification threshold and therefore the reported number of peptides identified with a given FDR. This fact complicates interpretation and comparison of results because database searches are usually not performed under the same conditions. In addition to the precursor mass tolerance, other factors that affect the number of candidates are the database taxonomy, the nature of the database, and the enzyme constriction set in the search engine. To study the extent to which these choices affect the accuracy of FDR estimation, we explored whether it was possible to estimate and even compensate for the impact of these effects on the FDR. Using SEQUEST, we searched the same population of MS/MS spectra analyzed in the previous section against a mouse DB (*reference search*), a *chordata* DB that included the mouse DB (*related search*), and a mouse DB doped with sequences from cyanobacteria proteins (*doped search*), using fully tryptic cleavages. We also searched against a mouse DB allowing semitryptic cleavages (*semitryptic search*). In all cases we used a 15 ppm precursor tolerance. We also searched the MS/MS spectra against separate decoy DBs generated from each of the target DBs. The false PSMs obtained in the different searches against the decoy DBs at given XCorr thresholds were then compared. As expected, and because of the increased number of candidate sequences, the related, doped, and semitryptic searches returned more incorrect hits than the reference search (Figure 2A–C). Consistently, when the FDR was calculated for each search condition using the separate DB method, the number of PSMs at a given FDR was lower in the related, doped, and semitryptic searches than in the reference search, reflecting a loss in sensitivity due to the increase in search space (Figure 2D–F).

We hypothesized that if the false assignments were truly random matches, it would be possible to control the search space effect by taking into account the relative random matching behavior of the different search conditions reflected in Figures 2A–C. We fitted these curves to linear functions, so that the number of decoy PSMs obtained at a given threshold score in the reference search (D_R) can be estimated from that obtained in any other search condition (D_L), and the decrease in false PSMs in the reference search can be estimated as $D_L - D_R$. Assuming that the number of true PSMs at the threshold XCorr score is not affected by changing these search conditions, the total number of peptides identified at the threshold (N_L) in a given condition would be expected to diminish by $D_L - D_R$ in the reference search. Therefore, from the FDR obtained in a search condition, given by $FDR_L = D_L/N_L$, it is possible to predict the FDR for the reference condition from $FDR_R = D_R/(N_L - D_L + D_R)$. Applying this procedure to the results obtained above, we found that the FDR reference curve estimated from the doped search (Figure 2E) and from the semitryptic search (Figure 2F) quite closely matched the experimentally determined FDR curve from the reference search and was only a slight overestimate when the calculation was made from the related search (Figure 2D). This slight overestimate was due to underestimation of the total number of PSMs in the reference search (compare Figure 2G with Figure 2H,I). Indeed, Venn diagram analysis revealed the existence of a significant number of identifications from *chordata* (related) proteins that were not contained in the population of mouse (reference) sequences, an effect not observed for overlap of identifications between the reference and the other two search conditions (compare Figure 2J with Figure 2K,L). A close inspection of the nature of the identified peptide sequences revealed that some of these matches came from closely related species, like rat, that contain highly homologous sequences, which by chance had the same mass and produced a slightly higher score than the mouse sequence (not shown). Disregarding these minor deviations, produced by the similarities between related species, our results demonstrate that the effect of the number of sequence candidates on the calculation of FDR can be well controlled by analyzing the random matching behavior in the decoy databases. Since these results were obtained using a precursor mass tolerance of 15 ppm, we conclude that when

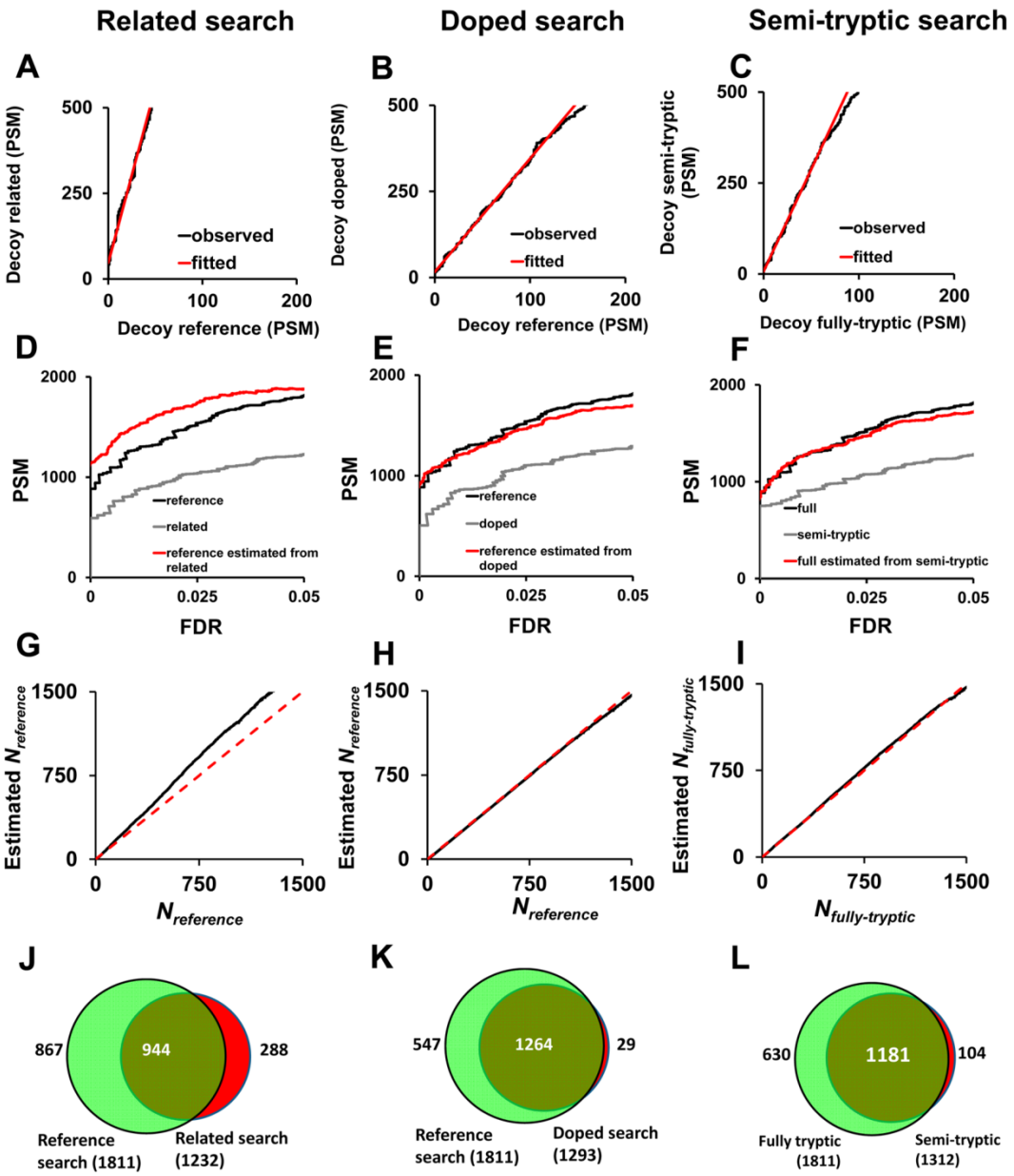


Figure 2. Estimation of FDRs in different searching conditions. The same set of MS/MS spectra from Figure 1 was subjected to fully tryptic search at 15 ppm precursor tolerance against a reference mouse DB, a larger related (*chordata*) DB, and a larger doped (mouse + cyanobacteria) DB, and to a semitryptic search against a mouse DB. Xcorr was used as score threshold for peptide identification. The number of PSM obtained in the different searching conditions was compared with those obtained in the reference search, as indicated (A–C, black lines). The red lines represent linear curve fittings of the data. The performance of peptide identification in each condition was compared with that of the reference condition as a function of FDR (D–F, black and gray lines). The number of identifications in the reference condition, estimated from the number obtained in each of the different conditions (as explained in the main text), is also represented (D–F, red lines). The estimated number of PSM in the reference search was compared with the observed number of PSM (G–I, black lines); the dashed red curves represent the identity lines. Data-proportional Venn diagrams of the sequences identified at FDR < 0.01% under each search condition are shown (J–L) together with the numbers of identified peptides. Note that >90% of the semitryptic-only PSM set (red region in L) contains semitryptic peptides.

independent scores like XCorr are used for PSM identification, estimation of FDR using decoy databases is a robust procedure that can be fully reproduced and controlled even when DB searches are done using narrow precursor mass tolerances.

Dispersion between Random Scores Produced by the Same MS/MS Spectra Increases as the Number of Candidate Peptide Sequences Decreases

The results presented in the previous sections demonstrate that the number of false PSMs can be accurately estimated by using a decoy DB; however, they do not provide information about the nature of the peptide sequences identified or the confounding effect of using DB-dependent scores. To further explore these factors, we inspected the joint distribution of SEQUEST XCorr scores obtained at wide (Figure 3A) or narrow (Figure 3B) mass precursor tolerances against the target and decoy forms of the NR DB, which, as explained above, contained sequences unrelated to the mouse DB and therefore represented false PSMs. These plots contained a cloud of points distributed symmetrically around

dx.doi.org/10.1021/pr5007284 | J. Proteome Res. 2015, 14, 700–710

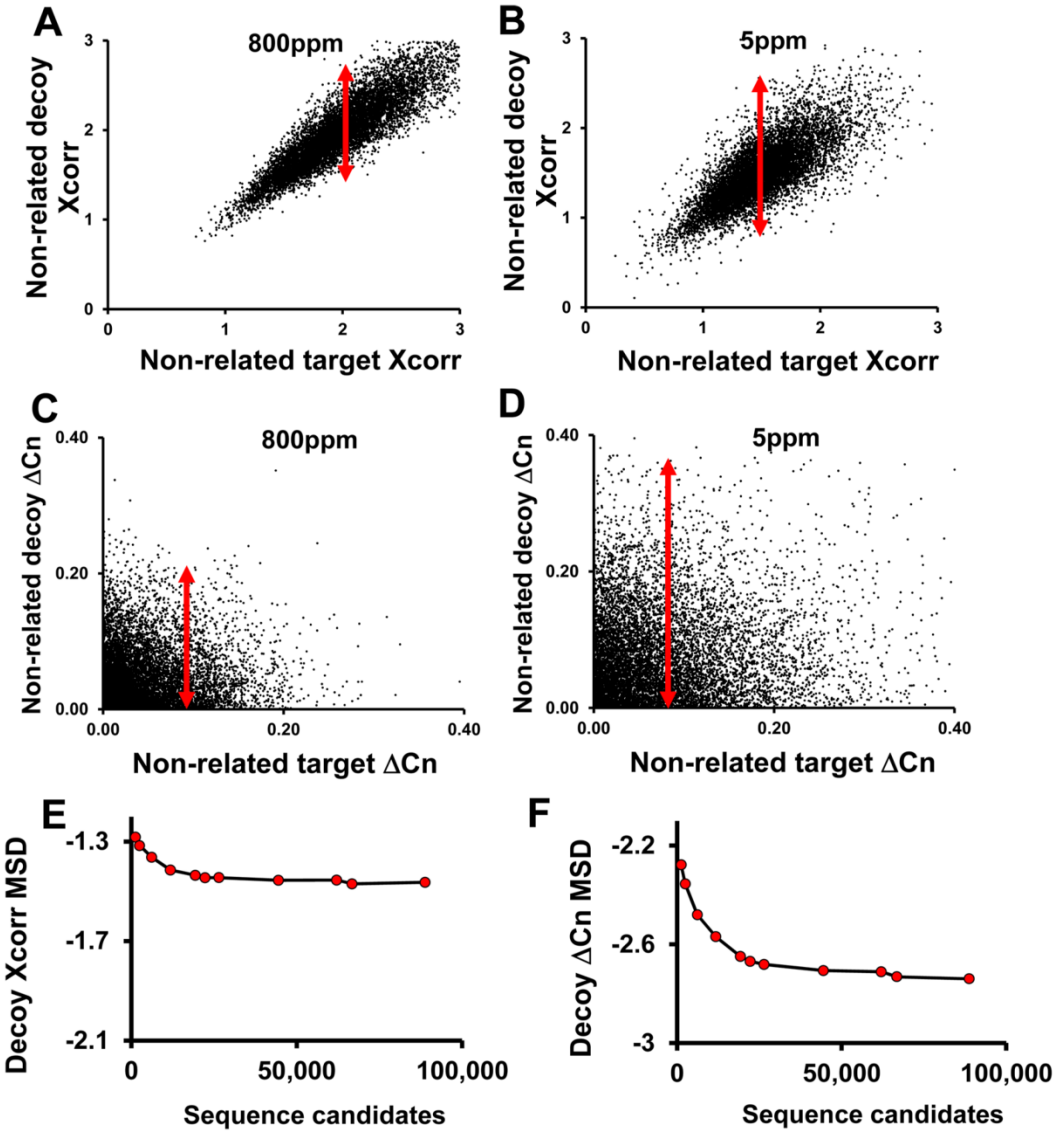


Figure 3. Effect of precursor mass tolerance on the joint distribution of SEQUEST scores between two decoy DBs. The collection of MS/MS spectra was searched separately against nonrelated (*cyanobacteria*) target or decoy DBs at 800 or 5 ppm, and the joint distribution of Xcorr (A) and ΔCn scores (B) were constructed. The mean-squared deviation of the scores around the diagonal line in panels A and B was then plotted against the number of sequence candidates (E,F).

the diagonal line; this symmetry explains why the number of PSMs obtained in the two DBs is identical at any score threshold, independently of the mass precursor tolerance. However, the dispersion of the cloud around the diagonal line in these plots is very marked, so that the agreement between decoy and target scores is actually rather poor, to the point that MS/MS spectra producing quite similar target scores produced very dissimilar decoy scores. For instance, the NR-decoy XCorr scores corresponding to NR-target scores around 2.0 fluctuated from less than 1.5 to more than 2.5 at 800 ppm (Figure 3A, red arrow). Interestingly, the dispersion of the cloud was more marked at 5 ppm; thus, for target XCorr scores around 1.5, the decoy scores ranged from less than 1.0 to more than 2.5 (Figure 3B). The same effect was observed in the joint distribution of SEQUEST ΔCn score; in this case, the dispersion of decoy scores corresponding to ΔCn target scores of ~0.1 at 5 ppm was almost double than that at 800 ppm (compare Figure 3C with Figure 3D). In addition, and in contrast with the results observed with XCorr, the joint distribution of ΔCn lacked the clear correlation between the scores obtained in the two DBs. To further analyze this effect, the search against the NR DBs was performed at a range of precursor mass tolerances, and the dispersion of the cloud, in terms of the mean-squared deviation of XCorr and ΔCn, was plotted against the average number of candidate peptide sequences per spectrum. We found that, as the number of candidates per spectrum increases, the dispersion of the two scores decreases, until a plateau is reached at around 20,000 sequence candidates (Figure 3E,F). These results show that, although the number of false PSMs obtained in a DB search can be accurately estimated, the random scores themselves are far from predictable, and the dispersion between the random scores obtained by the same MS/MS spectra increases when the search space decreases.

dx.doi.org/10.1021/pr5007284 | *J. Proteome Res.* 2015, 14, 700−710

Decreasing the Number of Sequence Candidates Can Significantly Alter the Nature of Identified Peptides When DB-Dependent Scores Are Used

To further explore the effect of decreasing the number of candidates in a real-world situation, we searched the set of 10,000 mouse MS/MS spectra against a target and a decoy mouse DB using precursor mass windows over the range from 3000 to 5 ppm; the nature of the peptides identified by XCorr alone (independent score) or by combining XCorr with ΔCn, using the probability ratio (pRatio) method[23] (DB-dependent score), were compared in the different conditions. Since it would be unrealistic not to take full advantage of the high mass accuracy provided by the mass spectrometer, in all the cases PSMs with a mass deviation above 5 ppm were filtered out[24] to remove false PSMs before FDR estimation. Since false PSMs are distributed over the entire mass precursor window, whereas true PSMs are expected to deviate less than 5 ppm from the experimental precursor mass, this 5 ppm-filtering step decreases the number of false PSMs and increases the performance of peptide identification.

The list of PSMs identified at 3000 ppm and postfiltered at 5 ppm (3000/5, corresponding to the results obtained using the highest number of sequence candidates) was selected as the reference, and the number of PSMs matching the same sequence in the same MS/MS spectrum as in the reference list and the percentage of nonmatching PSMs were both plotted against the number of candidate peptides per spectrum. As predicted, in both cases the number of PSM matching the reference list increased with the number of sequence candidates, reflecting how the low-ppm postfiltering step gradually increased peptide identification performance at wider precursor mass tolerances (Figure 4A,B). When the inference was made using XCorr alone, the vast majority of identifications at 5 ppm (1625) were contained within the list obtained at 3000/5 ppm (Figure 4E, in dark green). Moreover, the 419 PSM identified only at 3000/5 ppm, even though they had the same XCorr scores as those obtained at 5 ppm, were not considered true PSMs because they did not pass the FDR threshold at 5 ppm (Figure 4E, in light green). The percentage of PSM that was not contained in the reference list remained almost negligible (Figure 4C,E, in red). These results show that when an independent score is used alone, only the expected decrease in identification sensitivity is detected as the precursor mass window decreases, and no effect is observed on the nature of the identified peptides.

When the DB-dependent score pRatio was used, the identification yield at 5 ppm was 26% higher than that obtained using XCorr alone (2112 vs 1674 PSMs). While this increase in performance was expected, the DB-dependent score was also associated with a previously unobserved effect: the proportion of PSMs not contained in the reference list showed a clear increase as the number of candidates decreased, reaching almost 20% when the search was performed at 5 ppm (Figure 4D,F, in red). In these conditions, about one-third of the identified PSMs (644 out of 2415) were different between the two lists. From these, 303 PSMs were identified only at 3000/5 ppm due to the decrease in identification sensitivity, as above (Figure 4F, in light green); however, a surprising large amount of PSMs (341) were identified at 5 ppm but not at 3000/5 ppm (Figure 4F, in red). Most of these 341 PSMs had the same XCorr value as the PSMs obtained in the 3000/5 ppm search, but their ΔCn scores were considerably higher; as a consequence, they produced unexpectedly low pRatio values (i.e., better scores) at 5 ppm and therefore FDRs below the 1% threshold. These results highlighted an underlying unreliability of the DB-dependent score at these narrow mass precursor tolerances.

To study whether these results were extrapolatable to other situations, we repeated this analysis using several collections of MS/MS spectra produced from samples from very different backgrounds and using different DBs, and found the same effect in all cases studied. The percentage of PSM found not to coincide between results obtained at 5 and 3000/5 ppm ranged from 12% in a yeast experiment and 15% in a pig experiment to 26% in a human experiment. Finally, we analyzed with the same procedure a collection of MS/MS derived from the analysis of the UPS2 human protein standard mixture, detecting the same trend; in this case the discrepancy between the results obtained using narrow tolerances and the postfiltering approach amounted up to 29% of total PSM (Figure 4G,H). In this experiment, by counting up the number of PSM that did not belong to proteins from the UPS2 standard or expected contaminants it was possible to make a parallel estimate of the proportion of false assignations. In the case of the postfiltering approach, the proportion of false identifications remained below 2% using Xcorr and below 5% using pRatio (Figure 4G,H), indicating that the majority of PSMs were true assignments and that the target-decoy FDR estimates had a reasonable accuracy. In clear contrast, the proportion of false PSM in the population identified only at narrow tolerances (Figure 4H, red region) increased to an unacceptable rate of 15%, indicating that the quality of identifications was not being properly controlled. Taken together, all these results strongly suggest that the use of DB-dependent scores may produce increasingly unreliable results when the number of sequence candidates decreases.

Problems Associated with DB-Dependent Scores Affect Equally Both SEQUEST and Mascot Scores

To analyze in more detail the nature of the inconsistency of the results obtained by using pRatio at narrow windows and to determine whether this effect was also detected using other searching engines, we compared the results obtained at precursor mass windows of 5 and 800 ppm and after postfiltering by 5 ppm the results obtained at 800 ppm (800/5 ppm), by using both SEQUEST and Mascot in the same data set. Figure 5A–C illustrates how the decrease in the mass window from 800 to 5 ppm reduces the number of false positives (decoy PSM) above a given score; these figures also show how the 5 ppm postfiltering step eliminates the population of false positives outside the window, thus decreasing even more the number of decoy PSM (Figure 5C). Although a similar trend is found in the target PSM, as the score increases this population tends to concentrate the true positives, so that the results at 5 and 800/5 ppm tend to converge (Figure 5D). The net effect is that the same FDR is attained at a lower score threshold at 800/5 ppm (Figure 5E), thus increasing sensitivity of peptide identification (Figure 5F). The tendency of the 5 and 800/5 curves to converge as the fraction of true positives increases (Figure 5D) also explains why narrowing the precursor window only affects the sensitivity using Xcorr, but not the nature of the peptides identified. Taken together, these plots justify why the results obtained using Xcorr remain reliable even at low precursor tolerances.

Mascot[25] uses two parameters to evaluate the PSM, the ions score (IS) and the expectation value (EV). The IS is a cologarithmic transformation of the probability that the observed match between the MS/MS spectrum and the database sequence is a random event; therefore, IS only depends on the spectrum and the sequence and is independent from the other sequence

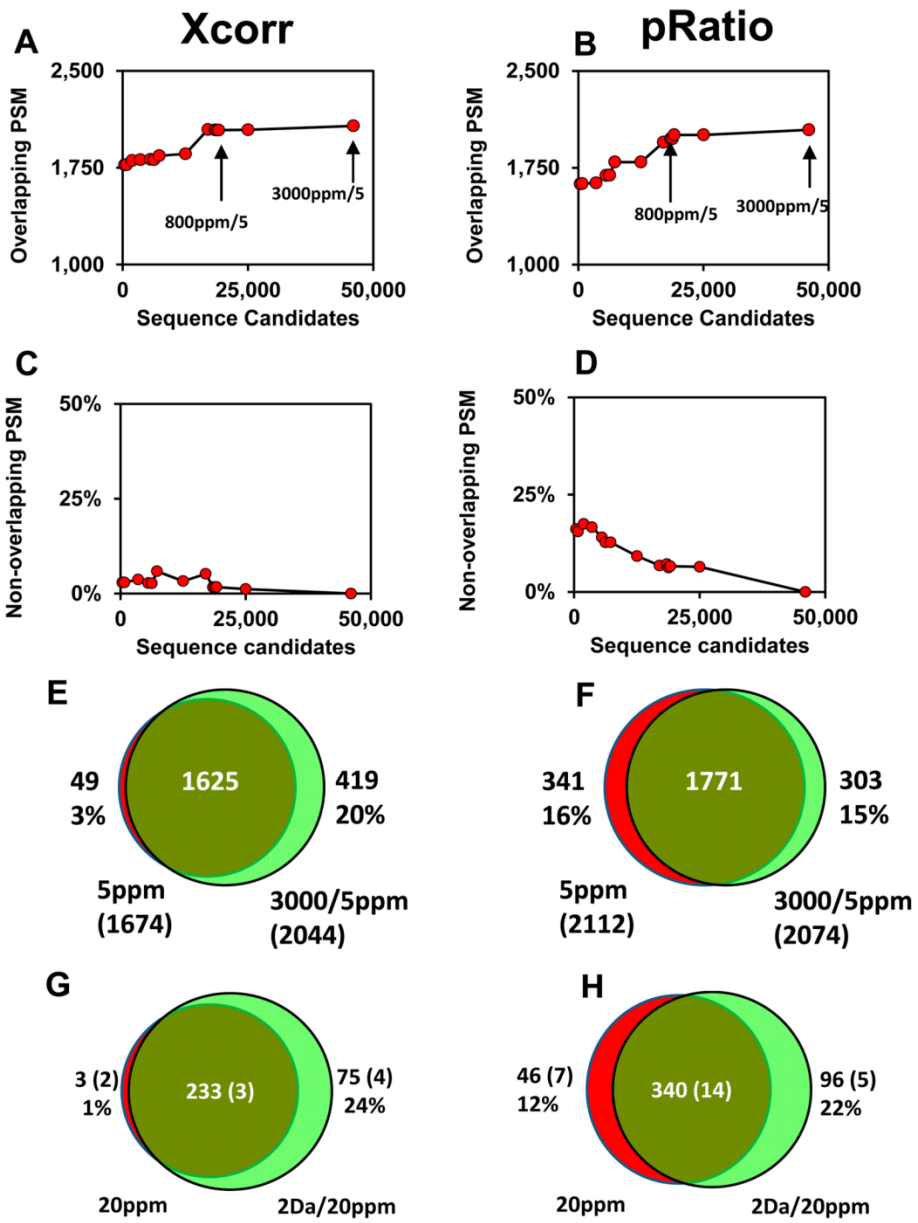


Figure 4. Effect of precursor mass tolerance on the nature of identified peptides. The collection of 10,000 MS/MS spectra was searched separately against a target or decoy mouse DB at different precursor tolerances, and XCorr alone or a combination of XCorr and ΔCn, using the probability ratio (pRatio) method, was used to analyze peptide identification. PSMs with a precursor mass error above 5 ppm were rescored to 0 in the case of XCorr or to 2 in the case of pRatio before FDR calculation. The number of PSM identified at 3000 pm followed by 5 ppm postfiltering that gave the same sequence match (overlap) than at the other precursor tolerances was plotted as a function of the average number of sequence candidates (A,B). The proportion of nonoverlapping PSMs was also analyzed (C,D). Data-proportional Venn diagrams showing PSM overlap between the two extreme conditions were also constructed for XCorr (E) or pRatio (F). The percentage of 5 ppm-only or 3000/5 ppm-only identified PSMs are given in parentheses. The same study was performed with the data produced from the analysis of a set of 48 human proteins (Sigma UPS2 reference mixture); in this case the MS/MS spectra were searched at 15 and 800 ppm precursor mass tolerances. The data-proportional Venn diagrams show PSM overlap between the results obtained at 5 ppm and 800/5 ppm for XCorr (G) or pRatio (H). The PSMs not belonging to the list of proteins in the standard or to known contaminantes were considered false identifications and are given in parentheses, together with the percentage of 5 ppm-only or 800/5 ppm-only identified PSMs.

candidates. Hence, IS must be considered an independent score and, as such, would be expected to have the same behavior as Xcorr with respect to the precursor mass window. Figure 5G–L shows how this prediction is fully confirmed, and although the differences between 5 and 800/5 ppm are less accused, there is a full resemblance between the behavior of Xcorr and IS. These data demonstrate that IS also produces reliable results at narrow precursor tolerances.

The same analysis was then performed using pRatio. Unlike Xcorr, pRatio is a DB-dependent score that depends on the second best score and therefore takes lower values (better scores) at lower precursor tolerances because the decrease in number of candidates increases the gap between scores (compare A and B in Figure 6). As a consequence, the number of both target and decoy PSM identified at a given score increases (Figure 6C,D). However, this effect is spectrum-specific; thus,

dx.doi.org/10.1021/pr5007284 | J. Proteome Res. 2015, 14, 700−710

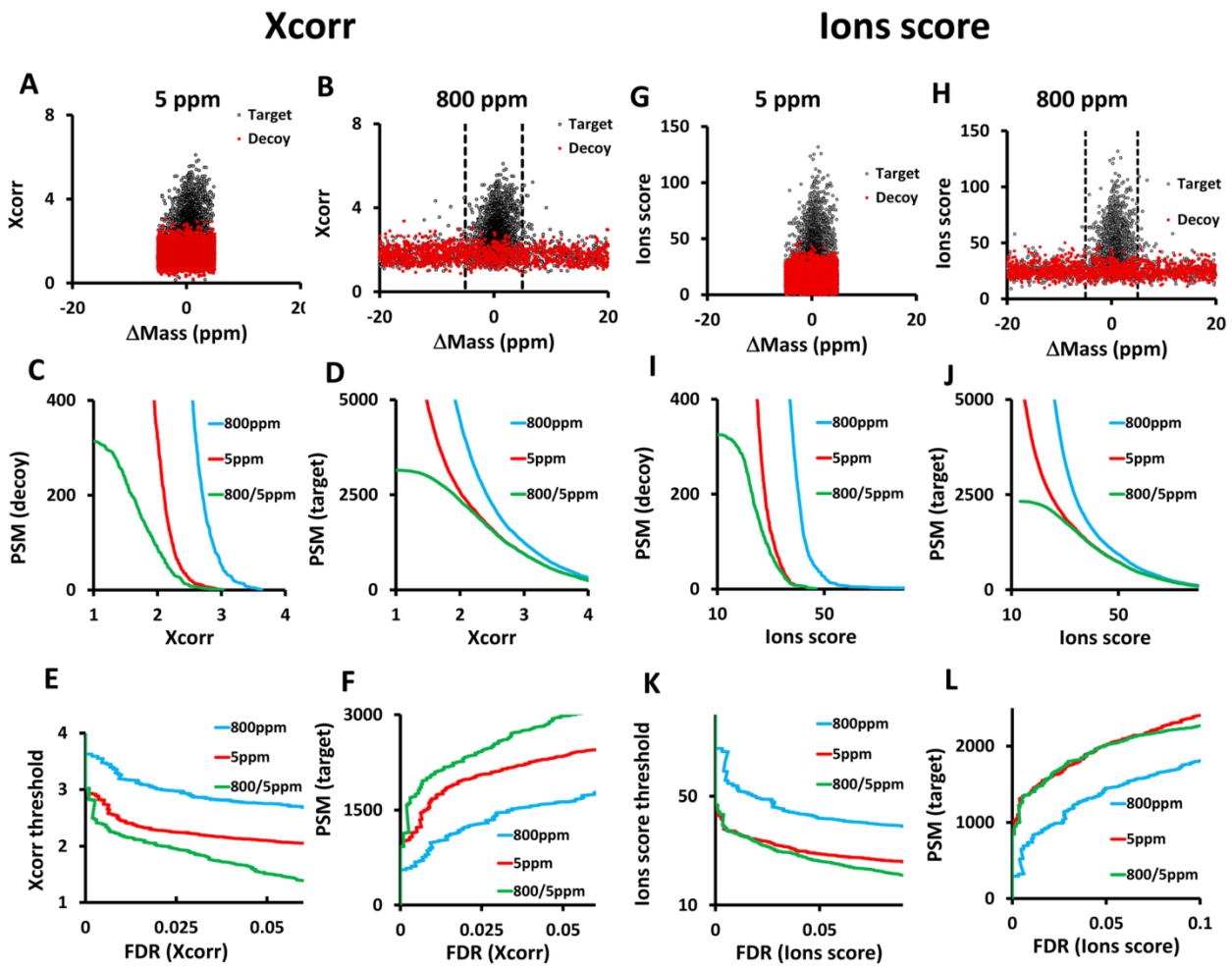


Figure 5. Comparative analysis of the behavior of the independent scores from SEQUEST (Xcorr) and Mascot (Ions score). The collection of 10,000 MS/MS spectra was searched separately against a target or decoy mouse DB at 5 and 800 ppm precursor tolerance, using SEQUEST and Mascot. Xcorr and Ions score, respectively, were used as scores for peptide identification. PSMs with a precursor mass error above 5 ppm were rescored to 0 before FDR calculation. The distribution of target and decoy hits was plotted as a function of mass deviation (A,B,G,H), showing how correct identifications distribute within a ±5 ppm window (marked by dashed lines). The effect of the narrow mass postfiltering on the number of decoy (C,I) and target PSM (D,J) was also represented as a function of Xcorr (C,D) or Ions score (I,J). The effect of the narrow mass postfiltering on the Xcorr (E) and Ions score (K) threshold needed to attain 1% FDR is presented. The performance of peptide identification using Xcorr (F) or Ions score (L) is also plotted as a function of FDR.

while the majority of spectra follow a linear trend when pRatio scores at 5 and 800/5 are compared in a cologarithmic representation (Figure 6G, green points), a significant proportion of points deviate from this trend, producing pRatio values at 5 ppm that surpass the score threshold and therefore are considered as positive PSM at 5 ppm but not at 800/5 ppm (Figure 6G, red points). This happens because these MS/MS spectra had low ΔCn values at 800/5 ppm but took anomalously high ΔCn values at 5 ppm when compared with the rest of the spectra (Figure 6H, red points), being therefore unreliable assignments. This population of anomalous matches, which correspond to the red regions of Venn diagrams in Figure 4F,H, demonstrates the instability of pRatio at narrow precursor tolerances.

The EV from Mascot depends on both the IS and the number of sequence candidates against which the spectrum is searched. Therefore, EV is also a DB-dependent score, and according to our line of reasoning, it should behave like pRatio. Analysis of the data, again, show a striking resemblance between these two scores (Figure 6I–N), confirming our predictions. In this case, the population of spectra that are positively matched at 5 ppm but not at 800/5 ppm is generated by a slightly different mechanism. These spectra suffer a more accused decrease in EV (increase in the cologarithm) at 5 ppm than the rest of the spectra (Figure 6O, red points) because they have an anomalously low number of sequence candidates at 5 ppm. Indeed, while the number of sequence candidates in this population had a distribution similar to the rest of the spectra at 800/5 ppm (Figure 6P, left graph); at 5 ppm the number of candidates clearly deviated from the trend followed by the rest of the spectra (Figure 6P, upper graph). Hence these positive matches are produced by the instability of EV due to a low number of sequence candidates.

■ CONCLUSIONS

In this article, we focus on a representative population of 10,000 spectra and analyze carefully the effect of search space on the accuracy with which the number of false PSMs is estimated using decoy databases. When we used independent scores, i.e., scores

dx.doi.org/10.1021/pr5007284 | J. Proteome Res. 2015, 14, 700−710

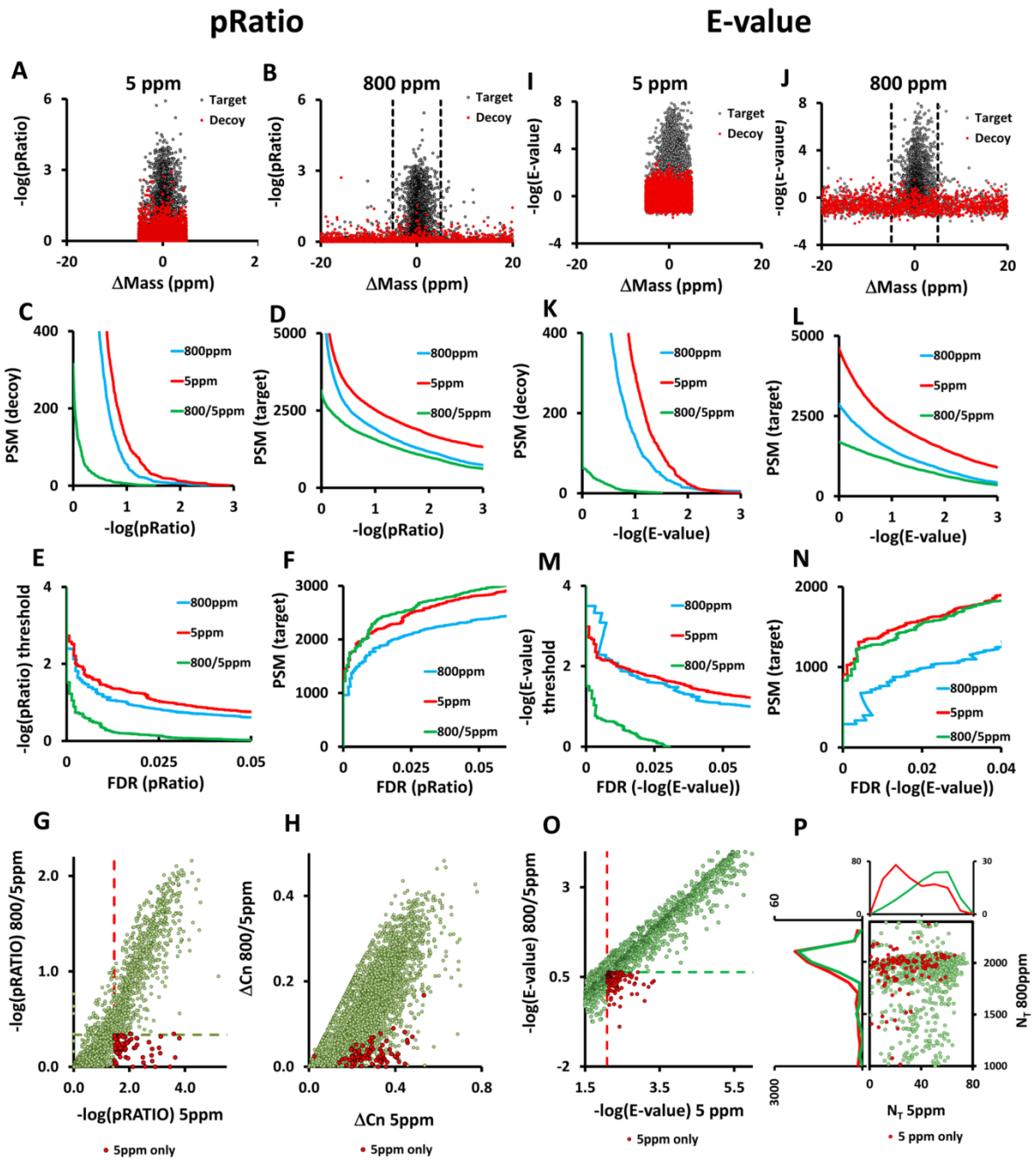


Figure 6. Comparative analysis of the behavior of the DB-dependent scores from SEQUEST (pRatio) and MASCOT (E-value). The upper graphs have the same meaning as those in Figure 5. In the postfiltering approach, PSMs with a precursor mass error above 5 ppm were rescored to 2 in the case of pRatio and to −10 in the case of −log(E-value) before FDR calculation. The results obtained at 800/5 were compared with those at 5 ppm, including −log(pRatio) (G), ΔCn (H), −log(E-value) (O), and the number of sequence candidates per spectrum (P). In all these plots, PSMs identified at 5 ppm and not identified at 800/5 ppm are depicted by red dots. Red and green dashed lines mark score thresholds at 5 and 800/5 ppm, respectively (G,O). In panel P, the frequency distribution of the number of sequences per spectrum is depicted in the top and left graphs; green and red lines represent the 800 and 5 ppm-only populations of PSM, respectively.

that only depend on the spectrum and the sequence that is matched, we were unable to detect a detrimental effect on this estimate even when searching against an average of only 200 sequence candidates. Indeed, estimates were so reproducible that we were able to make remarkably accurate predictions of FDR in diverse search conditions. However, we also found a marked difference in the behavior of DB-dependent scores, defined here as the scores that take additional information from other sequence candidates. We observed that the peptide sequences identified using DB-dependent scores when precursor mass tolerances are set in the low-ppm region can differ significantly from those obtained at wider tolerances, revealing an underlying inaccuracy in the estimation of FDR, which was due to a loss of reliability of DB-dependent scores when the number of sequence

dx.doi.org/10.1021/pr5007284 | J. Proteome Res. 2015, 14, 700−710

candidates decreases. This basic principle was demonstrated using two classic, well-known and still widely used searching engines, SEQUEST and Mascot, given that they both use independent and DB-dependent scores. However, our results may be easily extrapolated to other searching algorithms[26] once their scores are classified as being DB-dependent or independent. A relevant example is Andromeda,[27] a popular search engine that uses a DB-independent score similar to Mascot ions score and that, according to our line of reasoning, is not expected to lose reliability when very low precursor mass tolerances are used. We should note here that, although increasing the mass precursor window is expected to produce more reliable results with DB-dependent scores, it comes at the expense of considerable increases in search times. This is not only because of the increase in the number of sequence candidates that have to be scored but also because binning candidates by their mass, a common method to accelerate peptide search, will be less computationally effective. This factor should be taken into account for a proper choice between DB-dependent and independent scores.

The results presented here complement, from a different line of evidence, previous reports that detected a decrease in the discriminative power of delta scores, reflecting increased variability due to a significant reduction in the number of candidate peptide sequences.[6,15] On the basis of the data obtained in this work, we suggest that when DB-dependent scores are used, the high mass accuracy of modern mass spectrometers should not be exploited by filtering aggressively based on the precursor mass and then scoring peptide matches. Rather, and in line with the approaches used by an increasing number of authors,[6,13−16] we propose that more reliable results are to be expected when peptide matches are scored using wide precursor mass windows, avoiding inaccuracies derived from reduction of the search space, and then postfiltered according to their precursor mass.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: (+34) 91 4531200. Fax: (+34) 91 4531245. E-mail: jvazquez@cnic.es.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants BIO2012-37926 and ProteoRed-PT13/0001/0017 from Ministry of Economy and Competitiveness, and grant RD06/0014/0030 from Red de Investigación Cardiovascular (RIC, Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III, Ministry of Health). F.G-M. and M.T-H were supported by fellowships from the Ministry of Economy and Competitiveness. We thank S. Bartlett for English editing.

■ REFERENCES

(1) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. Neutron-encoded mass signatures for multiplexed proteome quantification. *Nat. Methods* **2013**, *10* (4), 332−4.

(2) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−14.

(3) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7* (1), 29−34.

(4) Navarro, P.; Vazquez, J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *J. Proteome Res.* **2009**, *8* (4), 1792−6.

(5) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−14.

(6) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73* (11), 2092−123.

(7) Cooper, B. The problem with peptide presumption and low Mascot scoring. *J. Proteome Res.* **2011**, *10* (3), 1432−5.

(8) Cooper, B. The problem with peptide presumption and the downfall of target-decoy false discovery rates. *Anal. Chem.* **2012**, *84* (22), 9663−7.

(9) Cottrell, J. S.; Creasy, D. M. Response to: the problem with peptide presumption and low Mascot scoring. *J. Proteome Res.* **2011**, *10* (11), 5272−3.

(10) Chalkley, R. J. When target-decoy false discovery rate estimations are inaccurate and how to spot instances. *J. Proteome Res.* **2013**, *12* (2), 1062−4.

(11) Navarro, P.; Trevisan-Herraz, M.; Bonzon-Kulichenko, E.; Núñez, E.; Martínez-Acedo, P.; Pérez-Hernández, D.; Jorge, I.; Mesa, R.; Calvo, E.; Carrascal, M.; Hernáez, M.; García, F.; Bárcena, J. A.; Ashman, K.; Abián, J.; Gil, C.; Redondo, J. M.; Vázquez, J. General statistical framework for quantitative proteomics by stable isotope labeling. *J. Proteome Res.* **2014**, *13* (3), 1234−1247.

(12) Jorge, I.; Navarro, P.; Martinez-Acedo, P.; Nunez, E.; Serrano, H.; Alfranca, A.; Redondo, J. M.; Vazquez, J. Statistical model to analyze quantitative proteomics data obtained by 18O/16O labeling and linear ion trap mass spectrometry: application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells. *Mol. Cell. Proteomics* **2009**, *8* (5), 1130−49.

(13) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285−92.

(14) Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. *Mol. Cell. Proteomics* **2008**, *7* (5), 962−70.

(15) Ding, Y.; Choi, H.; Nesvizhskii, A. I. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. *J. Proteome Res.* **2008**, *7* (11), 4878−89.

(16) Hsieh, E. J.; Hoopmann, M. R.; MacLean, B.; MacCoss, M. J. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* **2010**, *9* (2), 1138−43.

(17) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575−81.

(18) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582−7.
(19) Bonzon-Kulichenko, E.; Perez-Hernandez, D.; Nunez, E.; Martinez-Acedo, P.; Navarro, P.; Trevisan-Herraz, M.; Ramos Mdel, C.; Sierra, S.; Martinez-Martinez, S.; Ruiz-Meana, M.; Miro-Casas, E.; Garcia-Dorado, D.; Redondo, J. M.; Burgos, J. S.; Vazquez, J. A robust method for quantitative high-throughput analysis of proteomes by 18O labeling. *Mol. Cell. Proteomics* **2011**, *10* (1), M110 003335.
(20) Martinez-Bartolome, S.; Navarro, P.; Martin-Maroto, F.; Lopez-Ferrer, D.; Ramos-Fernandez, A.; Villar, M.; Garcia-Ruiz, J. P.; Vazquez, J. Properties of average score distributions of SEQUEST: the probability ratio method. *Mol. Cell. Proteomics* **2008**, *7* (6), 1135−45.
(21) Fitzgibbon, M.; Li, Q.; McIntosh, M. Modes of inference for evaluating the confidence of peptide identifications. *J. Proteome Res.* **2008**, *7* (1), 35−9.
(22) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−89.
(23) Martinez-Bartolome, S.; Navarro, P.; Martin-Maroto, F.; Lopez-Ferrer, D.; Ramos-Fernandez, A.; Villar, M.; Garcia-Ruiz, J. P.; Vazquez, J. Properties of average score distributions of SEQUEST: the probability ratio method. *Mol. Cell. Proteomics* **2008**, *7* (6), 1135−45.
(24) Hsieh, E. J.; Hoopmann, M. R.; MacLean, B.; MacCoss, M. J. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* **2010**, *9* (2), 1138−43.
(25) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−67.
(26) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73*, 2092−2123.
(27) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794−1805.

2. Second article

2.1 (English) General statistical framework for quantitative proteomics by stable isotope labelling

In this paper we present the WSPP statistical model (Weighted Spectrum, Peptide and Protein) (Navarro, 2014), a generally-applicable statistical framework for the analysis of data generated with SIL-MS (stable isotope labelling mass spectrometry) technologies. The model provides a detailed description of the behaviour of technical variance by analysing it independently at the spectrum, peptide and protein levels. Furthermore, the model is able to capture separately the specific error sources of each SIL and MS method, demonstrating that error distributions are accurately modelled in all cases at the three levels. It makes the integration at each level, taking into account the separate variances according to error propagation theory, so that the specific variance of each value at any level is accurately estimated. It resolves efficiently the undersampling problem, providing a framework to analyse the data at each level using unique normal distributions. In addition, the statistical framework also allows comparing and integrating results obtained using different SIL techniques, so that full control over variance is maintained in the integrated data, opening the possibility of making further integrations at higher levels. The WSPP model provides the first general framework for analysing quantitative SIL data on the basis of a unique and validated statistical model, using the Fundamental Workflow (the three-level workflow for data integrated from spectrum to peptide level, and from peptide to protein level).

The work presented in this paper has been conducted in close collaboration with Pedro Navarro and Elena Bonzon-Kulichenko, who are equally first authors. Pedro Navarro developed the statistical model previous to the WSPP, which was used specifically for ^{18}O-experiments using linear trap quadrupole (LTQ) mass spectrometres (Jorge, 2009; Bonzon-Kulichenko, 2011b). He was also a key contributor in the first stages of the generalisation of the model for other stable isotope labelling (SIL) techniques and other instruments, a work that has been presented in his PhD Thesis, along with QuiXoT, the associated software package. My contribution consists of the development of the subsequent stages of the algorithm and QuiXoT, initially in collaboration with him (for versions 1.2.66, May 2009, to version 1.3.11, February 2010), and in the last period on my own (myself being the only developer of QuiXoT since version 1.3.12, March 2010 to date); the development includes several satellite programs such as RAWToBinStack (since version 1.5, May 2010) and

QuiXtoQuiX (a software for averaging fullscan spectra, and other functions, developed between May 2009 and December 2012, and entirely developed by myself), among others. I am also the only developer of SanXoT (whose development started in August 2012), which implements the statistical model here presented, including the modifications explained in the third paper (see next section).

2.2 (Español) Un marco estadístico general para proteómica cuantitativa por marcaje isotópico estable

En este artículo presentamos el modelo estadístico WSPP (de sus siglas en inglés, *Weighted Spectrum, Peptide and Protein*) (Navarro, 2014), que proporciona un marco formal de aplicabilidad general para el análisis de datos generados por tecnologías de marcaje isotópico estable y espectrometría de masas (SIL-MS). El modelo describe en detalle el comportamiento de la varianza técnica al analizarla de manera independiente a nivel de espectro, péptido y proteína. Además, el modelo logra capturar por separado los errores específicos de cada método de marcaje y cada instrumento, demostrando que las distribuciones de error han sido modeladas con precisión en todos los casos a los tres niveles. Hace la integración a cada nivel teniendo en cuenta las varianzas por separado, de acuerdo con teoría de propagación de errores, de tal forma que la varianza específica de cada valor, en cada nivel, se puede calcular con precisión. Se resuelve eficientemente el problema de submuestreo (*undersampling*), aportando un marco de trabajo con el que se pueden analizar los datos a cada nivel usando distribuciones normales únicas. Adicionalmente, el marco estadístico permite comparar e integrar los resultados obtenidos con distintas técnicas de marcaje, de manera que el control total sobre la varianza se mantiene en los datos integrados, abriendo la posibilidad de hacer integraciones sucesivas a niveles superiores. El modelo WSPP da lugar al primer marco general para analizar datos cuantitativos por marcaje isotópico estable, fundamentado en un modelo estadístico único y validado, utilizando el Flujo de Trabajo Fundamental (la sucesión de tres niveles para datos integrados de espectro a péptido, y de péptido a proteína).

El trabajo presentado en este artículo se ha realizado en estrecha colaboración con Pedro Navarro y Elena Bonzon-Kulichenko, que, al igual que yo, también figuran como primer autor. Pedro Navarro desarrolló el modelo estadístico previo al WSPP, utilizado específicamente para experimentos de marcaje por ^{18}O, utilizando espectrómetros de masas de trampa lineal LTQ (*Linear Trap Quadrupole*) (Jorge, 2009; Bonzon-Kulichenko, 2011b). Él también ha sido un colaborador clave en las primeras fases de la generalización del modelo a otros métodos de marcaje isotópico y otros instrumentos, un trabajo que ha sido presentado en su tesis doctoral, junto con QuiXoT, el paquete de *software* asociado. Mi aportación consiste en los desarrollos posteriores del algoritmo y de QuiXoT, primero en colaboración con él (para las versiones entre 1.2.66, de mayo de 2009, hasta la versión 1.3.11, en febrero de 2010), y en el último periodo en solitario (siendo yo el único desarrollador de QuiXoT a partir de la versión 1.3.11, en febrero de 2010); el desarrollo comprende varios programas satélite, tales como RAWToBinStack (a partir de la versión 1.5, mayo de 2010), y

QuiXtoQuiX (programa que, entre otras funciones, promedia espectros *fullscan*, desarrollado entre mayo de 2009 y diciembre de 2012, y desarrollado por mi en su totalidad), entre otros. Yo soy también el único programador de SanXoT (iniciado en agosto de 2012), que implementa el modelo estadístico aquí presentado, añadiendo las modificaciones explicadas en el tercer artículo (véase siguiente apartado).

Journal of proteome research

Article

pubs.acs.org/jpr

General Statistical Framework for Quantitative Proteomics by Stable Isotope Labeling

Pedro Navarro,[†,‡,◆] Marco Trevisan-Herraz,[†,§,◆] Elena Bonzon-Kulichenko,[†,§,◆] Estefanía Núñez,[†,§] Pablo Martínez-Acedo,[†,§] Daniel Pérez-Hernández,[†,§] Inmaculada Jorge,[†,§] Raquel Mesa,[†,§] Enrique Calvo,[§] Montserrat Carrascal,[∥] María Luisa Hernáez,[⊥] Fernando García,[#] José Antonio Bárcena,[▽] Keith Ashman,[#,○] Joaquín Abian,[∥] Concha Gil,[⊥] Juan Miguel Redondo,[§] and Jesús Vázquez*[,†,§]

[†]Centro de Biología Molecular Severo Ochoa, CSIC−UAM, 28049 Madrid, Spain

[‡]Institute of Immunology, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany

[§]Centro Nacional de Investigaciones Cardiovasculares, 28029 Madrid, Spain

[∥]Instituto de Investigaciones Biomédicas de Barcelona (IIBB-CSIC), 08036 Barcelona, Spain

[⊥]Facultad de Farmacia, Universidad Complutense de Madrid, 28040 Madrid, Spain

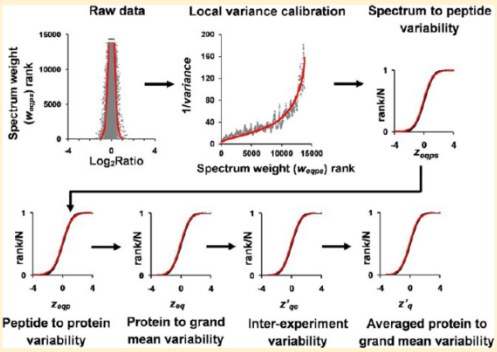
[#]Centro Nacional de Investigaciones Oncológicas, 28029 Madrid, Spain

[▽]Universidad de Córdoba and Córdoba Maimónides Institute for Biomedical Research (IMIBIC), 14004 Córdoba, Spain

[○]University of Queensland, Brisbane, St Lucia, Queensland 4072, Australia

Supporting Information

ABSTRACT: The combination of stable isotope labeling (SIL) with mass spectrometry (MS) allows comparison of the abundance of thousands of proteins in complex mixtures. However, interpretation of the large data sets generated by these techniques remains a challenge because appropriate statistical standards are lacking. Here, we present a generally applicable model that accurately explains the behavior of data obtained using current SIL approaches, including ^{18}O, iTRAQ, and SILAC labeling, and different MS instruments. The model decomposes the total technical variance into the spectral, peptide, and protein variance components, and its general validity was demonstrated by confronting 48 experimental distributions against 18 different null hypotheses. In addition to its general applicability, the performance of the algorithm was at least similar than that of other existing methods. The model also provides a general framework to integrate quantitative and error information fully, allowing a comparative analysis of the results obtained from different SIL experiments. The model was applied to the global analysis of protein alterations induced by low H_2O_2 concentrations in yeast, demonstrating the increased statistical power that may be achieved by rigorous data integration. Our results highlight the importance of establishing an adequate and validated statistical framework for the analysis of high-throughput data.



KEYWORDS: *Quantitative proteomics, stable isotope labeling, statistical analysis, yeast*

INTRODUCTION

Biological systems are dynamic and contain many molecular species (including DNA, RNA, proteins, carbohydrates, and lipids) whose interactions result in complex series of physicochemical, spatial, and temporal changes. Because proteins participate in all cellular processes, knowing how protein amounts and activities change over time provides important information about the state of the system. It is from such knowledge that the molecular mechanisms of disease and new pharmacological targets and biomarkers will emerge. Recent advances in mass spectrometry (MS)-based proteomics allow the identification and relative quantification of thousands of proteins in a single study. Despite these advances, the reproducibility of MS-based proteomics has been called into question.[1] It is an accepted fact that when the technology is properly applied, it is highly reproducible;[2] therefore, progress in the field will depend on a correct understanding of these techniques and their limitations.[3] The development of suitable statistical models is a critical step toward achieving this goal.

MS-based quantitative proteomics may be performed by direct quantitation of precursor or fragment ion peptide intensity in each of the samples (label-free approaches) or by using stable isotope labeling (SIL) techniques. In the most

Received: July 5, 2013
Published: January 31, 2014

dx.doi.org/10.1021/pr4006958 | J. Proteome Res. 2014, 13, 1234−1247

common setup, label-free quantitation involves analyzing several technical replicates of the same sample or samples from several subjects (biological replicates) belonging to two or more different conditions. Because of the multiplicative nature of the different factors (fixed effects) and error sources (random effects) involved, quantitative data are usually subjected to a logarithmic transformation that allows treating these effects as additive, providing a natural way for modeling the replicate structure of the data within the analysis of variance (ANOVA) framework.[4] A common feature of these ANOVA models is that the replicated structure are used to estimate the fixed effects and the variances associated with random errors, which are assumed to be normally distributed.[4−9] Although the analysis may be performed at the peptide feature level (one test per feature),[4,6] the quantitative results from different peptide features belonging to the same protein are usually integrated so that the analysis is at the protein level (one test per protein). Taking into account that the analysis is repeated for very large numbers of proteins (multiple hypothesis testing), statistically significant protein abundance changes under the different conditions are then detected by adjusting the *p*-value threshold to control for the false discovery rate (FDR).[10] In these models, all peptide features are considered to contribute equally to protein abundance, which may be estimated from the plain average of feature log-corrected intensities[4] or from features corrected by fixed effects[4,6,7] or scaled to the same level[9] before generating the protein average. In these approaches, random errors are assumed to derive from only one source so that the variance is a measure of the technical variability, but in some cases, biological variability is also taken into account by decomposing the total variance into the biological and the technical components.[4,6]

Stable isotope labeling (SIL) techniques, including stable isotope labeling with amino acids in cell culture (SILAC),[11,12] isobaric tagging for relative and absolute quantification (iTRAQ),[13] and enzymatic $^{16}O/^{18}O$ labeling,[14,15] currently offer the most accurate means of performing comparative quantitative proteomics studies.[16] Although the existence of separate technical, experimental,[4,6] and biological variations in iTRAQ analysis has been analyzed,[17] no statistical models were derived from these studies. The simplest model to analyze iTRAQ data is to calculate the protein value as an average of peptide ratios and to compare each of the protein values across several replicates (one test per protein) using an appropriate statistical test such as Student's *t* test. The protein average may be calculated using the mean,[18] the median,[19,20] or an average calculated by minimizing the square-root distance from the peptide readings from the log-transformed peptide ratios.[20] This kind of testing at the protein level has been extended using an ANOVA model, similar to those proposed to treat label-free data, with additive peptide effects and only one random effect that combines the biological error and the measurement noise,[21−24] which has been applied to the analysis of a case study involving four treatment groups with several replicates each.[8] All of these approaches have in common that all of the peptide readings originating from the same protein are equally considered, under the implicit assumption that they have the same variance. This assumption is based on original analysis showing that noncorrected ratios of peptides measured by iTRAQ follow approximately a log-normal distribution.[20] However, other analyses have demonstrated that not all of the peptides are quantified with the same accuracy, demonstrating a clear dependence of variance with ion intensity,[25−29] which may produce deviations from normality. Therefore, using intensity-weighted peptide averages of log ratios[26,30] to calculate protein averages has been proposed. Other approaches try to model or control the behavior of variance by a two-parameter modeling of the dependence of peptide variance with intensity[28] or by using a variance-stabilizing normalization (VSN) transformation[25,31] (similar to those employed in microarray approaches[32]), calculating the protein values as the median[31] or the trimmed average[25] of transformed peptide values. The transformed ratios at the spectrum level have been shown to have an apparently normal distribution, from which it is possible to detect statistically significant regulation of specific peptides (such as tyrosine-phosphorylated peptides) from the global distribution of transformed peptide ratios.[28] To test the significance of the changes at the protein level, most methods assume that the set of protein values follow a normal distribution and use a 0.05 probability threshold.[26,31] This test may be performed by direct fitting to a normal distribution[26] or by using estimates of the standard deviation,[31] whereas other approaches adjust the significance threshold so that 95% of experimental variation is encompassed.[25] Finally, some authors correct the significance value for multiple hypothesis testing.[31] Concerning SILAC data, the majority of studies use MaxQuant algorithm[33] to analyze quantitative results. In MaxQuant, protein ratios are calculated as the median of all SILAC peptide ratios, and the proteins are then grouped into bins according to their summed peptide intensities; in each bin, protein log -ratios are then assumed to be normally distributed, and the standard deviation is calculated using a robust estimate, from which a statistical significance is assigned to each protein. This procedure empirically takes into account the observed fact that high-abundance protein values have a lower variability than low-abundance ones.[33] Finally, we have recently proposed a statistical model to analyze quantitative data obtained by ^{18}O labeling, which decomposes the total technical variance into the spectrum, peptide, and protein variance components.[34] This approach models the heterogeneous variance at the spectrum level and integrates log ratios to the protein level using weighted averages according to error propagation theory. The validity of the model was demonstrated by showing that the distribution of protein values follows a normal distribution and by the very low percentage of outliers found at the spectrum, peptide, and protein levels.[34]

All of these studies show that, in spite of the efforts made in the field, a comprehensive statistical theory for the general analysis of quantitative data by SIL has not been developed yet. Existing models are highly specific to each SIL method and mass spectrometer, making them unsuitable for examining data from different laboratories, judging experimental quality on the basis of unified criteria, handling, comparing, and integrating multiple measurements, or interpreting the complete set of experimental results from different SIL approaches as a whole. Moreover, most models and statistical significance tests are based on normality assumptions that have not been tested despite the fact that heterogeneity of variance has been documented in all SIL methods.[25,33,34] These techniques are based on peptide-centric measurements, and the lack of general models leads to the subjective choice of a method for combining multiple peptide readings to estimate protein ratios.[25] This problem is further aggravated by the under-sampling that characterizes SIL-based MS analysis:[2] the number of peptides that quantify a protein is variable and cannot be

dx.doi.org/10.1021/pr4006958 | J. Proteome Res. 2014, 13, 1234−1247

Table 1. Statistical Parameters Estimated in the Null-Hypothesis Experiments

method	samples	weight constant (k_e)	spectrum variance (σ^2_S) (95% CI)	peptide variance (σ^2_P) (95% CI)	protein variance (σ^2_Q) (95% CI)	number of protein outliers (FDR_q < 0.05)
iTRAQ-TOFTOF	A vs A (114 vs 116)	182	0.029 (0.025−0.035)	0.012 (0.005−0.02)	0.002 (0−0.005)	0
iTRAQ-TOFTOF	B vs B (115 vs 117)	210	0.017 (0.013−0.021)	0.01 (0.005−0.015)	0.005 (0.002−0.009)	0
iTRAQ PQD	A vs A (114 vs 116)	14	0.110 (0.106−0.115)	0.010 (0.003−0.013)	0.003 (0.001−0.007)	1
iTRAQ PQD	B vs B (115 vs 117)	17	0.095 (0.091−0.100)	0.029 (0.022−0.037)	0.001 (0.000−0.005)	1
^{18}O-HR	A vs A	36	0.008 (0.0071−0.0085)	0.011 (0.009−0.012)	0.006 (0.0035−0.0064)	0
^{18}O-LR	A vs A	0.17	0.019 (0.018−0.021)	0.044 (0.038−0.050)	0.005 (0.002−0.009)	0
SILAC-HR	A* vs A	15	0.0001 (0.0001−0.0003)	0.0048 (0.004−0.005)	0.010 (0.009−0.012)	18
SILAC-LR	A* vs A	0.1	0.006 (0.0056−0.0063)	0.0082 (0.0069−0.0085)	0.0055 (0.005−0.0073)	9

Table 2. Statistical Parameters Estimated in the Control vs H_2O_2-Treatment Experiments

method	samples	weight constant (k_e)	spectrum variance (σ^2_S) (95% CI)	peptide variance (σ^2_P) (95% CI)	protein variance (σ^2_Q) (95% CI)	number of protein outliers (FDR_q < 0.05)
iTRAQ-TOFTOF	A vs B (114 vs 115)	236	0.018 (0.014−0.023)	0.0077 (0.003−0.013)	0.0269 (0.022−0.033)	1
iTRAQ-TOFTOF	A vs B (116 vs 117)	311	0.02 (0.014−0.028)	0.0073 (0.001−0.013)	0.0159 (0.010−0.022)	5
iTRAQ PQD	A vs B (114 vs 115)	15	0.11 (0.106−0.114)	0.017(0.012−0.021)	0.0058 (0.001−0.010)	12
iTRAQ PQD	A vs B (116 vs 117)	16	0.11 (0.105−0.113)	0.024 (0.019−0.028)	0.009 (0.005−0.015)	6
^{18}O-HR	A vs B	32	0.0076 (0.0060−0.0080)	0.042 (0.039−0.046)	0.035 (0.030−0.041)	12
^{18}O-LR	A vs B	0.17	0.003 (0.0025−0.0032)	0.021 (0.020−0.023)	0.011 (0.009−0.015)	8
SILAC-HR	A* vs B	9.4	0.0004 (0.0002−0.0006)	0.0037 (0.003−0.004)	0.018 (0.017−0.02)	39
SILAC-LR	A* vs B	0.16	0.0041 (0.0039−0.0043)	0.007 (0.0065−0.0074)	0.021 (0.019−0.023)	14

controlled between experiments, a nontrivial fact that complicates mathematical modeling.

Here, we describe the WSPP (weighted spectrum, peptide, and protein) model, a generally applicable statistical framework for the analysis of data generated with SIL-MS technologies. The model provides a detailed description of the behavior of technical variance, and by analyzing it independently at the spectrum, peptide, and protein levels, the model is able to capture separately the specific error sources of each SIL and MS method, demonstrating that error distributions are accurately modeled in all cases at the three levels. The model generates the integration at each level taking into account the separate variances according to error propagation theory so that the specific variance of each value at any level is accurately estimated. The model efficiently resolves the under-sampling problem, providing a framework to analyze the data at each level using unique normal distributions. In addition, the statistical framework also allows comparing and integrating results obtained using different SIL techniques so that full control over variance is maintained in the integrated data, opening the possibility of making further integrations at upper levels. The WSPP model provides a general framework for analyzing quantitative SIL data on the basis of a unique and validated statistical model.

■ MATERIALS AND METHODS

Yeast Culture, SILAC Labeling, and Sample Preparation

Cells from the lysine auxotrophic *Saccharomyces cerevisiae* strain YMJ38 (his3Δ1 leu2Δ0 ura3Δ0 arg4Δ::kanMX4 trp1Δ::kanMx4 lys2Δ) were cultured and SILAC-labeled with [$^{13}C_6$]L-lysine and [$^{13}C_6$]L-arginine as previously described.[35] SILAC light (A) and heavy (A*) media were prepared from minimal synthetic dextrose medium (0.5% (w/v) ammonium sulfate, 2% (w/v) glucose, 0.17% (w/v) nitrogen base without amino acids, and 20 mg/L of amino acids except for lysine and arginine).[35] Cells were grown in the corresponding medium at 30 °C and with shaking at 200 rpm until OD_{600} = 0.8. After adjusting both cell cultures to OD_{600} = 0.4 with the corresponding SILAC medium, culture A was split in two equal volumes, 20% (v/v) H_2O_2 (Panreac) was added to one of them (sample B) at a final concentration of 0.5 mM, and the other sample remained as a control (sample A). The three cell cultures (A, A*, and B) were grown at 30 °C with shaking at 200 rpm for 3 h until OD_{600} = 1.6 and were then centrifuged at 3000*g* at 4 °C for 10 min. Cell pellets were washed three times with cold distilled water, taken up in 500 μL of homogenization buffer (50 mM Tris-HCl, pH 7.5, 10% glycerol, 1% Triton X-100, 150 mM NaCl, 0.1% SDS, and 5 mM EDTA) supplemented with 1 mM PMSF and the protease inhibitors cocktail from Sigma-Aldrich (St. Louis, MO) and lysed using a homogenizer with glass Ballotini beads (0.45 μm) (Sigma). Cell lysates were centrifuged (13 000*g*) at 4 °C for 15 min, and protein concentration in the supernatant was determined by Bradford assay (Biorad). The cleared lysates were stored in aliquots of 100 or 500 μg of protein.

Experimental Design of Quantitative Techniques

The lysate aliquots from the three samples (A, non treated; B, treated; and A*, nontreated, SILAC-labeled) were subjected to high-throughput quantitation using the iTRAQ, ^{18}O, and SILAC-labeling techniques using both low- and high-resolution instruments in a total of 16 experiments (Tables 1 and 2) performed in different laboratories. Eight experiments were performed to test the null hypothesis associated with each technique (Table 1), and another eight experiments were performed to analyze the effect of H_2O_2 treatment (Table 2). Note that because of the use of four-plex iTRAQ reagents we could perform two replicate comparisons in the same experiment.

Protein Digestion and Labeling and Peptide Fractionation

In each experiment, the lysate aliquots were independently digested, and the resulting peptides were labeled (except in the SILAC cases) so that all of the replicates included all of the

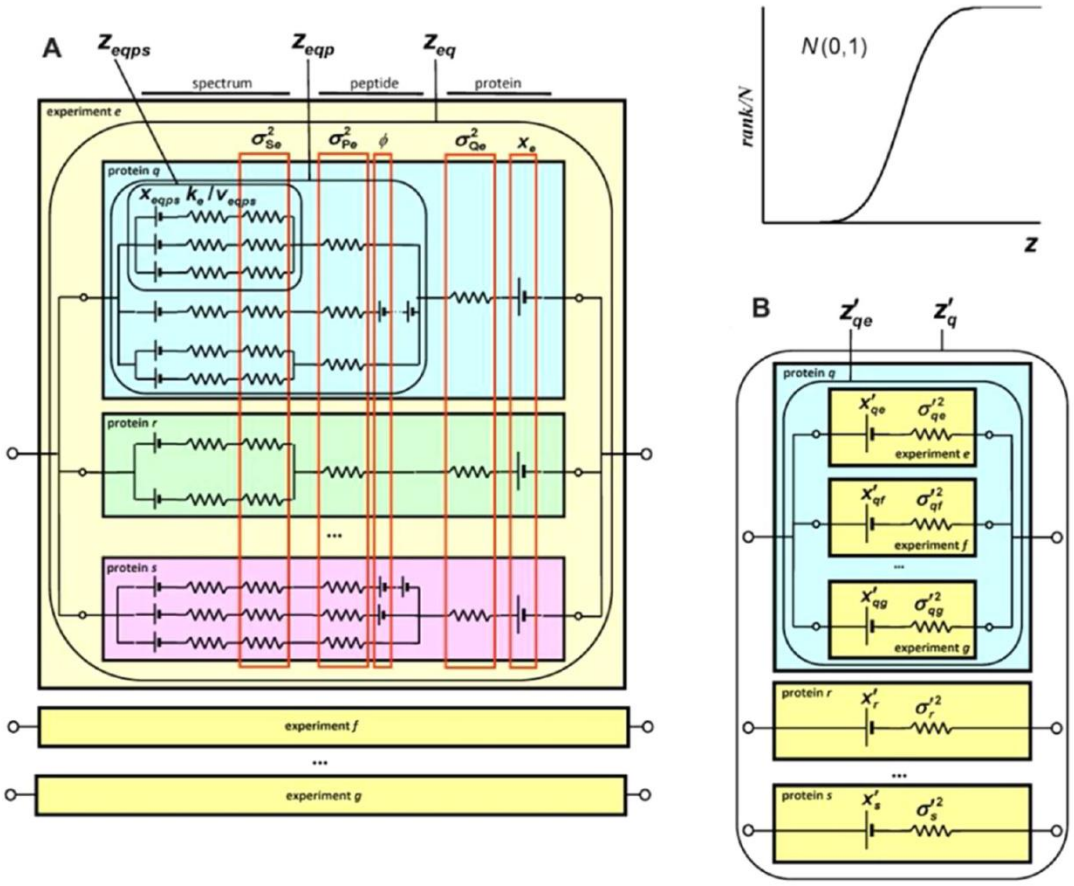


Figure 1. Representation of the WSPP model as an equivalent electric circuit. (A) In a quantitative SIL experiment, each protein is quantified by one or more peptides and each peptide is quantified in one or more spectra, represented by the branches of an electrical circuit. Resistances and batteries correspond to the variances and $\log_2$ ratios of each element, respectively. Within a given experiment, spectrum branches contain a spectrum-specific resistance (k_e/v_{eqps}) and a constant resistance (σ^2_{Se}), whereas peptide and protein branches have constant resistances (σ^2_{Pe} and σ^2_{Qe}). The correction for Arg to Pro conversion in SILAC experiments, estimated by a maximum likelihood method, is equivalent to a battery (ϕ) for each Pro residue in the peptide branches, and the systematic experiment error is equivalent to a constant battery (x_e) in the protein branches. (B) Integration of technical replicates is equivalent to setting the protein circuits from different experiments in parallel. In the null hypothesis, normal distributions with zero mean and unit variance are followed by the seven standardized parameters describing the variability in units of standard deviation among spectra within a peptide (z_{eqps}), among peptides within a protein (z_{eqp}), among proteins within a experiment (z_{eq}), among experiments within a protein (z'_{qe}), and among proteins (z'_q) (inset).

technical error sources. Each lysate aliquot was trypsin-digested separately using the whole proteome in-gel digestion protocol we described previously.[15] Prior to digestion, the gel bands were reduced with 10 mM DTT and alkylated with 50 mM iodoacetamide.[45] iTRAQ labeling was performed essentially according to the manufacturer's instructions; details of the procedure are described in the Supporting Information. ^{18}O labeling was performed following the robust protocol that we have previously described in detail.[15] In the experiments in Table 2, sample B was labeled with ^{18}O and sample A was labeled with ^{16}O. In all of the experiments and prior to MS analysis, peptides were IEF-separated into 24 fractions exactly as described previously.[15]

Mass Spectrometry Analysis

Low-resolution analysis of SILAC- and ^{18}O-labeled peptides was performed using a linear ion trap LTQ (Thermo-Finnigan) in the Cardiovascular Proteomics Laboratory at the Centro de Biología Molecular Severo Ochoa (Madrid). High-resolution analyses of SILAC- and ^{18}O-labeled peptides were performed using an LTQ-Orbitrap XL ETD (Thermo-Finnigan) in the Proteomics Unit of the Centro Nacional de Investigaciones Cardiovasculares (Madrid). iTRAQ-labeled samples were analyzed using either a linear ion trap LTQ (Thermo-Finnigan) working in the PQD scanning mode in the Proteomics Facility of the IDIBAPS (Barcelona) or a MALDI-TOF/TOF (Applied Biosystems) in the Proteomics Unit of the Centro Nacional de Investigaciones Oncológicas (Madrid). Details of the methods used are described in the Supporting Information.

Peptide Identification

Peptide identification was performed using either SEQUEST or Mascot. SEQUEST results were analyzed using the probability ratio method,[36] taking into account isoelectric points of peptides to improve peptide identification.[15] In all cases, false discovery rates (FDR) of peptide identifications were calculated from the search results against the inverted databases using the refined method.[15] Full details are described in the Supporting Information. The identification data were deposited at the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[37] under data set identifier PXD000325.

dx.doi.org/10.1021/pr4006958 | J. Proteome Res. 2014, 13, 1234−1247

Peptide Quantification and Statistics

For low-resolution MS data (LTQ) obtained by ^{18}O labeling, peptide quantification from ZoomScan spectra and calculation of labeling efficiencies of peptides were performed by fitting the spectra to a theoretical function as described previously[38,39] using QuiXoT, a program written in C# in our laboratory, which is available at http://www.cnic.es/en/inflamacion/proteomica/wiki/. Information about the meaning of the parameters used in QuiXoT is also available in the Supporting Information. For high-resolution MS data (LTQ-Orbitrap) obtained by ^{18}O labeling, quantification was performed directly in the full-scan spectrum with the same algorithm used for calculating labeling efficiency by taking into account the theoretical isotopic envelope of the identified peptide,[39] with the exception that mass peaks were not fitted to a Gaussian/double-exponential distribution but to the maximum intensity of each MS peak within a 0.02 m/z window around the theoretical m/z value. The same methods were used for low- and high-resolution SILAC data, respectively, except that the difference between the labeled and nonlabeled species was calculated from the peptide sequence and no correction for labeling efficiency was used. For iTRAQ data, only the intensity of the reporter ions within 0.4 Da windows around the theoretical values was taken into account for the quantification. Reporter intensities are corrected for isotopic contaminants by taking the information provided by the manufacturer into consideration. For each spectrum, a fitting weight was calculated upon quantification according to Supporting Information Table 2 by taking into account the intensity of peaks and the mean squared deviation between the theoretical curve and the experimental data. The data analysis workflow was performed as previously described.[34] The statistical model used to analyze the quantitative data is described in full detail in the Supporting Information and is schematized in Figure 1. Briefly, the $\log_2$ ratio of concentration in the two samples being compared, A and B, determined by spectrum s of peptide p derived from protein q in experiment e is expressed as $x_{eqps} = \log_2(A/B)$. The $\log_2$-ratio value associated with each peptide, x_{eqp}, is then calculated as a weighted average of the spectra used to quantify the peptide, and the value associated with each protein, x_{eq}, is similarly the weighted average of its peptides. In addition, a grand mean, x_e, is calculated in each experiment as a weighted average of the protein values. The statistical weights in all cases are the inverse of the local variances of each of the spectrum, peptide, and protein values. The local variance of each spectrum is modeled as a hyperbolic function of its fitting weight using two parameters, k_e and σ^2_{Se}. The local variances of the peptide and protein values are then calculated by error propagation theory by taking into account two additional constant variances: at the peptide, σ^2_{Pe}, and at the protein level, σ^2_{Qe}. The four constant parameters, k_e, σ^2_{Se}, σ^2_{Pe}, and σ^2_{Qe}, which describe the error distribution in each experiment, are estimated from all of the data together by an iterative method that uses robust approaches. The global distribution of values at each one of the levels is described using a standardized variable, z, that expresses the quantitative values in units of standard deviation and that in the null hypothesis is expected to follow a $N(0,1)$ normal distribution. Outliers at the scan and peptide levels and significant protein-abundance changes are detected from the z values by using a false discovery rate (FDR) threshold of 5%.

RESULTS

WSPP: A Statistical Model with a Multilayered Structure That Can Be Described by an Equivalent Electrical Circuit

The WSPP model separately considers the variances produced during (i) protein extraction and manipulation, (ii) peptide generation from their corresponding proteins and labeling, and (iii) generation of quantitative information from the mass spectra. We reasoned that this three-layered structure was the most appropriate for addressing the diverse sources of error generated by different labeling and MS approaches. The complete mathematical formulation is described in the Supporting Information, and the parameters used by the model are described in Supporting Information Table 1. The WSPP model takes into account under-sampling of data and uses robust algorithms to estimate variances. The model can be represented by a mathematically equivalent electrical circuit (Figure 1A) in which variances at the spectral, peptide, and protein levels, being independent, are therefore additive and can be represented as a series of resistances. Similarly, when replicates at the spectral or peptide levels are integrated to produce peptide or protein averages, their variances are treated following the same rule used for resistances located in parallel. Voltages represent abundance ratios in $\log_2$ scale. Branches with lower resistance contribute more to the final voltage, mirroring the fact that measurements with lower variance, which are more accurate, are weighted more in the averages. The equivalent circuit also shows how the addition of more branches at a given level decreases the resistance, reflecting the fact that the error of quantification diminishes when more values are averaged. However, the constant resistances set in series, which represents the constant error sources at each level, put a lower limit to this effect. This property reflects the fact that extensive averaging of a large number of spectra will never eliminate the error made at the peptide level, and averaging large number of peptides does not eliminate the error made at the protein level. Hence, averages are not affected much by very intense peaks or proteins with a very large number of peptides, as often happens in other weighting schemes.

To test the theory, we used a model system consisting of protein extracts from *S. cerevisae* cultures: sample A was obtained from untreated cells, sample A*, from untreated cells labeled by SILAC, and sample B, from cells treated with 0.5 mM H_2O_2. To ensure that data came from technical replicates, protein samples were prepared only once and stored in aliquots, and each aliquot was processed separately. The aliquots were digested with trypsin, and the resulting peptides were labeled (except for SILAC) and fractionated by isoelectric focusing using a robust protocol developed previously.[15] To analyze the null hypothesis (NH), A was compared with A or B was compared with B; for SILAC samples, pseudonull hypotheses were analyzed instead by comparing A* with A. To analyze the effect of H_2O_2 treatment, samples A and B were compared either by iTRAQ or ^{18}O labeling, and SILAC A* was compared with B. A total of 32 sample aliquots were processed for a total of 16 pairwise quantitative experiments (Tables 1 and 2). Peptide fractions were distributed among five laboratories and were analyzed on different mass spectrometers. SILAC and ^{18}O samples were analyzed using both low-resolution LTQ linear ion traps and a high-resolution hybrid LTQ-Orbitrap, whereas iTRAQ data were analyzed using both low-energy PQD fragmentation on an LTQ and high-energy

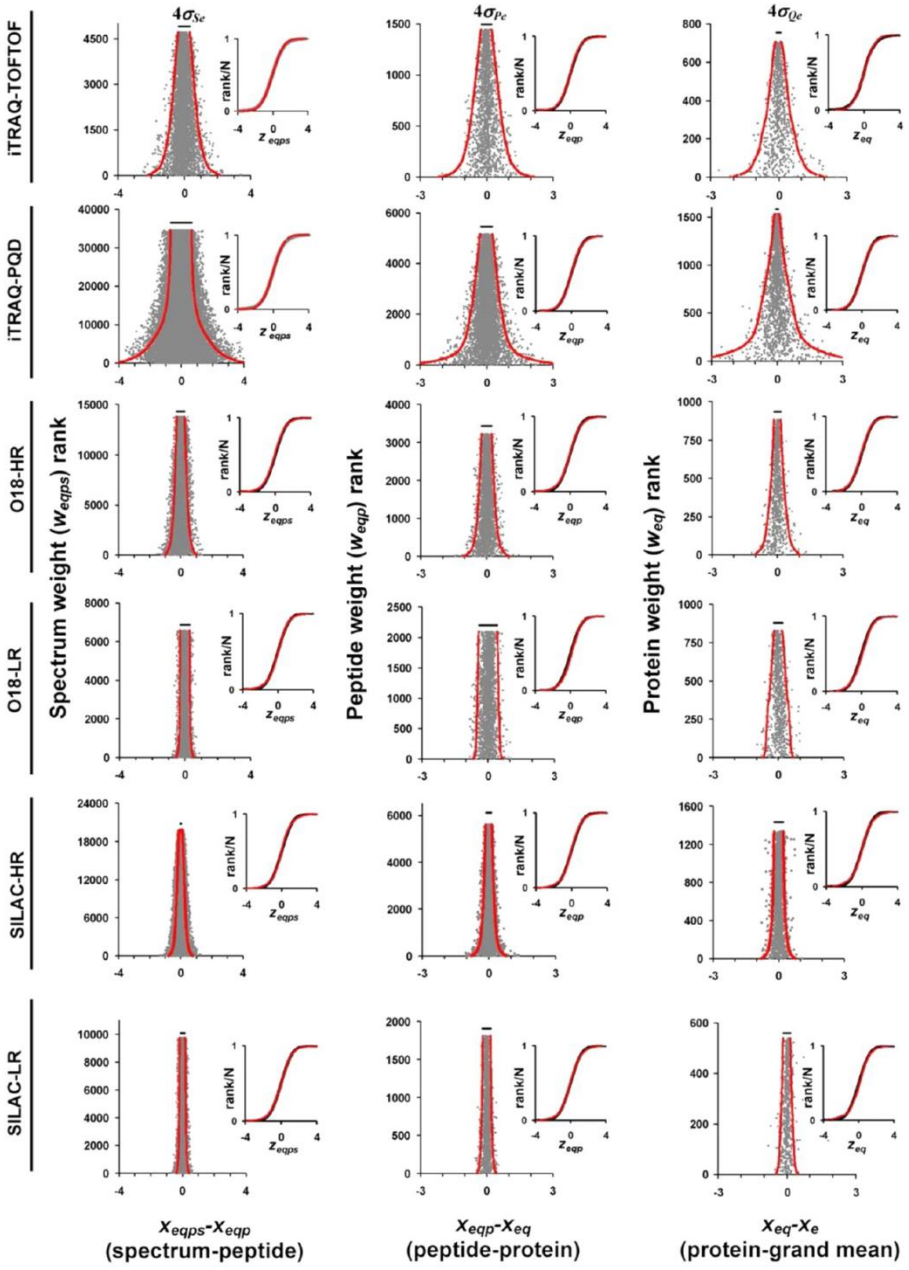


Figure 2. WSPP model gives a very accurate description of the distribution of quantitative errors under the null hypothesis at the spectral, peptide, and protein levels for all of the SIL methods analyzed. The plots show, for each one of the six SIL methods as indicated, the weight distributions of (left panels) $\log_2$ ratios of individual quantifications (spectra) around the corresponding peptides, (central panels) peptides around their corresponding proteins, and (right panels) proteins around the grand mean of the experiment. For better clarity, in the iTRAQ cases, only one of the two possible comparisons is shown. Red lines indicate local 95% confidence intervals (two standard deviations) in each case, as predicted by the model; horizontal black bars indicate the minimum (asymptotic) interval value. Insets show the cumulative distributions of the standardized variables described in Figure 1 (black points) at each level and SIL method; red lines are drawn according to the theoretical normal distribution with zero mean and unit variance, highlighting the excellent agreement between results and theory. In the true null-hypothesis experiments, only one false abundance change at the protein level (in iTRAQ-PQD) was detected as statistically significant at a 5% FDR among more than 1500 proteins (Tables 1 and 2).

CID fragmentation on a MALDI-TOF/TOF. The resulting MS data from all procedures were then gathered and analyzed.

Errors at the Spectrum, Peptide, and Protein Levels Can Be Modeled by Normal Distributions

Quantitative SIL experiments are peptide-centric approaches where ions are directly quantified in the MS detector; because not all spectra produce equally precise quantifications, each spectrum has a different variance and hence quantitative data cannot, in general, be treated as a whole by using a unique normal distribution.[34] The dependence of variance with ion intensity has been reported in several works.[25−29] The same trend was observed in the quantitative experiments performed to test the null hypothesis in this work; as shown in Figure 2,

dx.doi.org/10.1021/pr4006958 | J. Proteome Res. 2014, 13, 1234−1247

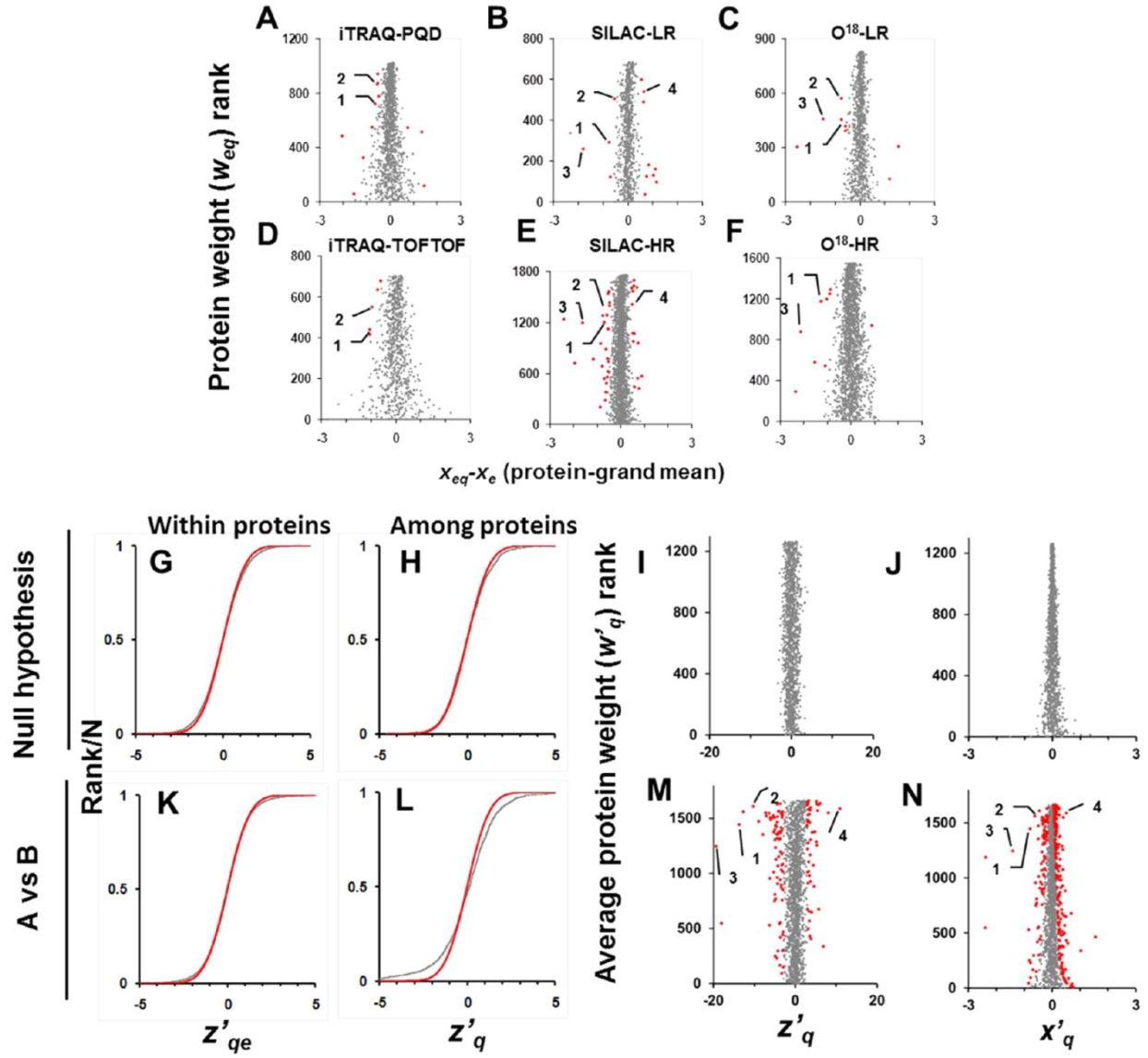


Figure 3. Detection of statistically significant protein-abundance changes produced by H_2O_2 treatment. (A–F) Weight distributions of protein quantifications around the grand mean obtained by each of the indicated SIL methods. In these plots, negative values signify an increase in protein abundance (toward the left) and positive values, a decrease (toward the right). Outliers at the protein level ($FDR_{eq} < 0.05$) are highlighted in red. Numbers indicate protein outliers that are consistently detected by most SIL methods: 1, cytochrome c peroxidase; 2, thioredoxin-2; 3, glutathione peroxidase 2; and 4, alcohol dehydrogenase 4. (G–L) Distribution of quantitative errors within and among proteins after integration of results from all SIL experiments. Upper panels correspond to the null-hypothesis experiments and lower panels, to the control vs H_2O_2-treatment study. The sigmoid plots show the distributions of the standardized variables describing variability of results from different experiments within the same protein (z'_{qe}, G, K) or among the different proteins (z'_q, H, L). The right panels show the weight distributions of z'_q (tornado plots I, M) and x'_q(J, N); red points indicate proteins showing a significant abundance change ($FDR'_q < 0.05$).

left column, the dispersion of quantitative data was not homogeneous but decreased with the statistical weight, a parameter that depended on ion intensity (see below). These data confirmed the heterogeneity of variance in all SIL and MS combinations analyzed.

The WSPP model addresses this issue by calculating a fitting weight (v_{eqps}) for each spectrum; this parameter considers quality features, such as peak intensity or goodness of fit, that are characteristic of the MS detector and SIL method used. The fitting weights were carefully designed (Supporting Information Table 2) to allow ranking by increasing accuracy of the entire collection of spectra in a given experiment so that quantifications from spectra having the same weight locally follow a normal distribution in all of the cases (Supporting Information Figure 1). The collection of fitting weights in a given experiment are then considered together to estimate the calibration constant (k_e) and the asymptotic spectrum variance (σ^2_{Se}) (Supporting Information Figure 2). These two parameters, which are characteristic of each experiment, are used to estimate separately the local variance of each one of the spectra from the fitting weights ($k_e/v_{eqps} + \sigma^2_{Se}$, Figure 1A). To study the distribution of $\log_2$ ratios at the spectrum level as a whole, we calculated a standardized variable, which expresses the $\log_2$-ratio deviation of the spectrum from the peptide average in

dx.doi.org/10.1021/pr4006958 | J. Proteome Res. 2014, 13, 1234–1247

Table 3. Integration of Protein-Abundance Changes in the Control vs Treated Experiments

Protein Description[a]	Corrected $\log_2$-ratio: 18O HR	18O LR	iTRAQ PQD 114-115	iTRAQ PQD 116-117	iTRAQ TOFTOF 114-115	iTRAQ TOFTOF 116-117	SILAC HR	SILAC LR	x'_q	Fold change		z_{eq}: 18O HR	18O LR	iTRAQ PQD 114-115	iTRAQ PQD 116-117	iTRAQ TOFTOF 114-115	iTRAQ TOFTOF 116-117	SILAC HR	SILAC LR	z'_q	FDR'_q	Comments
sp|Q04120|TSA2 Peroxiredoxin TSA2	-3.09	-2.53	-2.05	-1.77			-2.42	-2.32	-2.38	5.21	up	-13.93	-13.41	-8.91	-7.83			-16.78	-14.14	-31.23	4.E-211	oxidative stress response, regulated by Yap1
sp|P38143|GPX2 Glutathione peroxidase	-2.13	-1.49	-0.59	-0.70			-1.61	-1.79	-1.39	2.63	up	-8.70	-9.54	-3.17	-3.43			-11.14	-10.39	-19.33	1.E-80	oxidative stress response, regulated by Yap1 and Skn7
sp|P41816|OYE3 NADPH dehydrogenase 3	-3.43						-1.94		-2.39	5.25	up	-14.21						-12.23		-18.03	3.E-70	oxidative stress response, regulated by Yap1
sp|P00431|CCPR Cytochrome c peroxidas	-1.26	-0.76	-0.63	-0.61	-0.74	-0.97	-0.77	-0.76	-0.78	1.72	up	-5.81	-4.85	-4.10	-3.63	-5.01	-5.76	-5.34	-4.48	-13.63	5.E-40	oxidative stress response, regulated by Yap1
sp|P22803|TRX2 Thioredoxin-2	-0.59	-0.75	-0.55	-0.73	-0.58	-0.78	-0.47	-0.54	-0.62	1.53	up	-2.62	-5.42	-4.54	-5.14	-5.28	-5.53	-3.29	-3.55	-12.68	1.E-34	oxidative stress response, regulated by Yap1
sp|P02365|RS6 40S ribosomal protein S	-0.25	-0.54	-0.52	-0.35	-0.48	-0.62	-0.25	-0.28	-0.45	1.37	up	-1.12	-3.41	-4.18	-2.69	-5.65	-5.93	-1.75	-1.88	-10.20	3.E-22	oxidative stress response, regulated by Skn7
sp|P29509|TRXB1 Thioredoxin reductase	-0.88	-0.58	-0.48	-0.29	-0.56	-0.47	-0.40	-0.44	-0.49	1.40	up	-4.19	-3.66	-3.47	-1.82	-3.80	-2.64	-2.91	-2.66	-8.85	1.E-16	oxidative stress response, regulated by Yap1 and Skn7
sp|P54124|ALDH3 Aldehyde dehydrogenas	-1.02	-0.60	-0.74	-0.46	-0.26	-0.54	-0.50	-0.35	-0.50	1.41	up	-4.73	-3.17	-3.52	-1.50	-3.02	-3.38	-3.56	-1.93	-7.88	3.E-13	oxidative stress response
sp|P05743|RL26A 60S ribosomal protein	-0.12	-0.44	-0.53	-0.35	-0.36	-0.26	-0.23	-0.24	-0.35	1.28	up	-0.58	-2.66	-5.04	-2.69	-2.95	-1.85	-1.59	-1.54	-7.16	8.E-11	ribosomal protein
sp|P38616|YGP1 Protein YGP1	-0.42	-0.42	-0.22	-0.38	-0.41	-0.59	-0.07	-0.38	-0.36	1.28	up	-1.63	-2.66	-1.85	-2.61	-3.57	-4.48	-0.46	-2.19	-7.01	2.E-10	oxidative stress response, regulated by Msn2/4
sp|P02406|RL28 60S ribosomal protein	-0.56	-0.16	-0.15	-0.38	-0.43	-0.28	-0.30	-0.31	-0.31	1.24	up	-2.64	-1.14	-1.09	-2.24	-4.12	-2.29	-2.14	-2.03	-6.34	2.E-08	oxidative stress response, regulated by Yap1
sp|P53829|CAF40 Protein CAF40	0.05						-1.17		-0.84	1.79	up	0.18						-7.42		-6.24	3.E-08	CCR4-NOT core complex, regulation of transcription
sp|Q96VH4|HBN1 Putative nitroreductas			-0.15	-0.03	-1.24	-1.07		-0.42	-0.59	1.51	up			-0.72	-0.12	-6.25	-4.56		-1.98	-6.17	4.E-08	oxidative stress response
sp|P05756|RS13 40S ribosomal protein	-0.09	-0.19	-0.37	-0.42	-0.35	-0.43	-0.25	-0.29	-0.32	1.24	up	-0.40	-1.27	-2.42	-2.47	-3.15	-3.08	-1.77	-1.86	-6.03	1.E-07	ribosomal protein
sp|P26781|RS11 40S ribosomal protein	-0.24	-0.42	-0.24	-0.48	-0.34	-0.33	-0.27	-0.26	-0.32	1.25	up	-1.11	-2.75	-2.16	-3.49	-1.88	-1.50	-1.89	-1.65	-5.93	2.E-07	ribosomal protein
sp|P05753|RS4 40S ribosomal protein S	-0.14	-0.25	-0.29	-0.33			-0.32	-0.31	-0.28	1.22	up	-0.68	-2.05	-2.96	-2.70			-2.28	-2.09	-5.43	3.E-06	ribosomal protein
sp|P09938|RIR2 Ribonucleoside-diphosp	-0.05	-0.33	-0.18	-0.18	-0.31	-0.61	-0.23	-0.25	-0.27	1.21	up	-0.25	-2.65	-1.50	-1.16	-2.21	-3.64	-1.67	-1.66	-5.28	6.E-06	DNA replication
sp|P05150|OTC Ornithine carbamoyltran	-0.17	-0.34	-0.71	-0.40	-0.01	-0.83	-0.70		-0.47	1.39	up	-0.75	-1.64	-2.73	-1.24	-0.01	1.19	-4.80		-5.21	9.E-06	arginine synthesis
sp|P38061|RL32 60S ribosomal protein	-0.26	-0.43	-0.26	-0.39			-0.28	-0.31	-0.32	1.25	up	-0.98	-2.68	-2.21	-2.81			-1.55	-1.79	-5.18	1.E-05	ribosomal protein
sp|P14126|RL3 60S ribosomal protein L	0.25	-0.26	-0.30	-0.34	-0.32	-0.02	-0.32	-0.23	-0.26	1.20	up	1.19	1.98	-3.14	-2.89	-0.86	-0.04	-2.32	-1.55	-5.00	2.E-05	ribosomal protein
sp|P33315|TKT2 Transketolase 2	-1.08		-1.14	-0.64			-0.28		-0.57	1.48	up	-3.66		-3.79	-1.96			-1.90		-4.99	2.E-05	oxidative stress response, regulated by Msn2/4
sp|Q03102|YMN1 Uncharacterized membra			-0.77	-0.46	-1.17	-0.30	-0.37		-0.54	1.45	up			-3.73	-1.51	-2.76	-0.50	-2.58		-4.98	2.E-05	oxidative stress response via the HOG pathway
sp|Q03104|MSC1 Meiotic sister chromat	-0.63	-0.31	-0.13	-0.12			-0.32	-0.42	-0.30	1.23	up	-3.08	-2.54	-0.99	-0.79			-2.32	-2.79	-4.96	3.E-05	enviornmental stress response
sp|P39938|RS26A 40S ribosomal protein	-0.36	-0.22	-0.25	0.03	-0.43	-1.03	-0.31	-0.47	-0.33	1.26	up	-1.19	-1.36	-1.38	0.15	-1.75	-3.73	-2.14	-2.89	-4.91	3.E-05	ribosomal protein
sp|P42846|KRI1 Protein KRI1	-2.35						-0.30		-0.81	1.75	up	-7.06						-1.56		-4.87	4.E-05	ribosome biogenesis
sp|P05738|RL9A 60S ribosomal protein	-0.03	-0.28	-0.25	-0.60	-0.03	-0.32	-0.26	-0.30	-0.27	1.21	up	-0.13	-2.26	-1.66	-3.57	-0.14	-1.34	-1.83	-1.97	-4.78	6.E-05	ribosomal protein
sp|P05754|RS8 40S ribosomal protein S	0.00	-0.29	-0.23	-0.43	-0.26	-0.35	-0.29	-0.30	-0.28	1.21	up	0.00	-2.32	-1.51	-2.41	-1.35	-1.40	-2.04	-1.99	-4.77	6.E-05	ribosomal protein
sp|P06115|CATT Catalase T	-0.25	-0.44	-0.52	-0.37	-0.29	-0.42	-0.36	-0.28	-0.35	1.27	up	-1.17	-2.75	-2.47	-1.15	-0.71	0.85	-2.56	-1.57	-4.77	5.E-05	oxidative stress response, regulated by Msn2/4
sp|Q07653|HBT1 Protein HBT1	-0.97						-0.50	-0.44	-0.54	1.45	up	-3.06						-3.07	-2.47	-4.77	5.E-05	oxidative stress response, regulated by Msn4
sp|Q12068|GRE2 NADPH-dependent methyl	-0.90	-0.60	-0.47	0.01				0.15	-0.37	1.30	up	-4.27	-3.66	-2.99	0.04				0.83	-4.61	1.E-04	oxidative stress response via the HOG pathway
sp|P22943|HSP12 12 kDa heat shock pro	-0.03	-0.34	-0.14	-0.36			-0.40	-0.43	-0.29	1.22	up	-0.13	-1.63	-1.16	-2.50			-2.78	-2.71	-4.56	1.E-04	enviornmental stress response
sp|P05735|RL19 60S ribosomal protein	-0.35	-0.21	-0.29	-0.37	0.08	0.35	-0.31	-0.27	-0.26	1.20	up	-1.67	-1.64	-2.18	-2.37	0.23	0.91	-2.21	-1.79	-4.52	2.E-04	ribosomal protein
sp|P14796|RL40 60S ribosomal protein			-0.77	-0.76					-0.77	1.70	up			-3.37	-2.96					-4.49	2.E-04	ribosomal protein
sp|P22768|ASSY Argininosuccinate synt	0.30	-0.11	-0.17	-0.43			-0.40	-0.40	-0.25	1.19	up	1.39	-0.78	-1.64	-3.34			-2.84	-2.67	-4.41	3.E-04	oxidative stress response, regulated by Skn7
sp|P05030|PMA1 Plasma membrane ATPase	0.12	-0.13	-0.11	-0.25	-0.19	-0.23	-0.29	-0.25	-0.18	1.13	up	0.54	-1.10	-1.18	-2.20	-2.05	-2.02	-2.12	-1.62	-4.35	3.E-04	oxidative stress response
sp|P47137|YJ66 Uncharacterized oxidor	0.64		-0.21	0.17			-0.35	-0.53	0.41	1.33	up	2.83		-0.56	0.43			-2.41	-2.93	-4.34	3.E-04	oxidative stress response
sp|P32827|RS23 40S ribosomal protein	-0.18	-0.47	-0.28	-0.23			-0.20	-0.26	-0.29	1.22	up	-0.61	-3.12	-2.14	-1.45			-1.31	-1.53	-4.33	3.E-04	ribosomal protein
sp|P05740|RL17A 60S ribosomal protein	-0.44	-0.16	-0.15	-0.27	-0.17	-0.34	-0.21	-0.21	-0.22	1.17	up	-2.03	-1.22	-1.26	-1.89	-1.22	-2.06	-1.47	-1.40	-4.27	4.E-04	ribosomal protein
sp|P14065|GCY Protein GCY	-0.72	-0.43	-0.53	0.10					-0.43	1.35	up	-3.15	-2.86	-2.43	0.37					-4.27	4.E-04	oxidative stress response
sp|P39078|TCPD T-complex protein 1 su	-0.24		-0.66	-1.06			-0.15	-0.39	-0.36	1.29	up	-1.12		-2.66	-3.54			-1.04	-2.36	-4.20	0.001	protein folding
sp|Q03246|RT17 37S ribosomal protein	-0.39						-0.69		-0.65	1.57	up	-0.86						-4.10		-4.14	0.001	ribosomal protein
sp|P00447|SODM Superoxide dismutase [	-0.68	-0.23	-0.20	-0.20	-0.16	-0.08	-0.29	-0.26	-0.26	1.20	up	-2.97	-1.70	-1.08	-0.96	-0.73	-0.29	-2.04	-1.67	-4.13	0.001	oxidative stress response
sp|P00445|SODC Superoxide dismutase [	-0.15	-0.40	-0.24	0.00	-0.16	-0.54	-0.02	-0.03	-0.20	1.15	up	-0.68	-2.69	-2.25	-0.03	-1.36	-3.82	-0.16	-0.20	-4.04	0.001	oxidative stress response, regulated by Yap1
sp|P47771|ALDH2 Aldehyde dehydrogenas	-0.34	-0.71	-0.53	-0.33	-0.04	-0.19	-0.14		-0.32	1.25	up	-1.70	-2.86	-2.77	-1.46	-0.13	-0.53	-1.01		-3.99	0.001	oxidative stress response
sp|P34227|PRX1 Mitochondrial peroxire	-0.60	-0.42	-0.17	-0.20	-0.11	-0.09	-0.15	-0.08	-0.25	1.19	up	-3.34	-3.05	-1.13	-1.06	-0.40	-0.09	-1.06	-0.51	-3.96	0.001	oxidative stress response
sp|P33442|RS3A 40S ribosomal protein	-0.11	-0.29	-0.19	-0.20	-0.05	-0.29	-0.20	-0.14	-0.19	1.14	up	-0.52	-2.26	-1.70	-1.53	-0.42	-1.99	-1.43	-0.90	-3.87	0.002	ribosomal protein
sp|P32337|IMB3 Importin subunit beta-	0.01						-0.59		-0.51	1.42	up	0.02						-4.14		-3.84	0.002	ribosome biogenesis
sp|P23248|RS3B 40S ribosomal protein	-0.39		-0.33	-0.16	-0.13	-0.40	-0.16	-0.36	-0.26	1.19	up	-1.54		-1.68	-0.73	-0.91	-2.30	-1.06	-2.13	-3.83	0.002	ribosomal protein
sp|P39741|RL35 60S ribosomal protein	-0.60	0.04	-0.35	-0.03			-0.29	-0.34	-0.25	1.19	up	-2.70	0.24	-2.60	-0.19			-1.95	-2.06	-3.79	0.003	ribosomal protein
sp|P38962|TVP23 Golgi apparatus membr	0.20						0.87		0.69	1.61	up	0.44						4.24		3.67	0.004	oxidative stress response, regulated by Msn2/4
sp|P36148|GPT2 Glycerol-3-phosphate O					-0.13	-0.55	-0.63		-0.48	1.40	up					-0.51	-1.83	-3.51		-3.66	0.004	phospholipid biosynthesis
sp|Q07896|NOC3 Nucleolar complex-asso	-0.36						-0.55		-0.52	1.44	up	-0.99						-3.53		-3.63	0.004	ribosome biogenesis
sp|P38013|AHP1 Peroxiredoxin type-2 O	-0.31	-0.22	-0.14	-0.19	-0.17	-0.34	0.16	0.20	-0.15	1.11	up	-1.40	-1.81	-1.44	-1.55	-2.04	-3.29	1.12	1.35	-3.62	0.005	oxidative stress response via the HOG pathway
sp|P49334|TOM22 Mitochondrial import		-0.18	-0.10	-0.15			0.62		-0.35	1.27	up		-1.20	-0.34	-0.47			-4.06		-3.60	0.005	mitochondrial membrane organization
sp|Q04432|HSP31 Probable chaperone pr	-0.61	-0.29	-0.12	-0.14	-0.32	-0.02			-0.24	1.18	up	-2.95	-1.84	-0.73	-0.75	-2.51	-0.10			-3.59	0.005	oxidative stress response
sp|P04449|RL24A 60S ribosomal protein	-0.15	-0.22	-0.19	-0.23			-0.28	-0.24	-0.22	1.17	up	-0.70	-1.44	-1.47	-1.46			-1.95	-1.56	-3.57	0.005	ribosomal protein
sp|P16547|OM45 Mitochondrial outer me	-0.22	-0.50	-0.46	-0.15	0.08	1.34	-0.32	-0.12	-0.27	1.20	up	-0.92	-2.36	-2.54	-0.72	0.15	2.14	-2.08	-0.73	-3.51	0.007	mitochondrial membrane organization
sp|Q02753|RL21A 60S ribosomal protein	0.22	-0.15	-0.18	-0.33			-0.28	-0.33	-0.22	1.16	up	1.00	-0.80	-1.51	-2.36			-1.51	-2.12	-3.48	0.007	ribosomal protein
sp|P49631|RL43 60S ribosomal protein	-0.12	-0.29	0.02	-0.31	-0.37	-0.35	-0.12	-0.25	-0.20	1.15	up	-0.41	-1.53	0.18	-1.95	2.21	-1.81	-0.84	-1.54	-3.46	0.008	ribosomal protein
sp|P53315|SOL4 Probable 6-phosphogluc	-1.07	-0.22			-0.83	0.24	-0.29		-0.35	1.27	up	-3.07	-1.34			-2.07	0.43	-1.89		-3.42	0.009	pentose phosphate pathway
sp|P06367|RS14A 40S ribosomal protein	-0.18	-0.28	-0.04	-0.09	-0.33	-0.11	-0.12	-0.19	-0.17	1.12	up	-0.79	-2.13	-0.34	-0.60	-2.71	-0.72	-0.83	-1.28	-3.39	0.009	ribosomal protein
sp|P04451|RL23 60S ribosomal protein	-0.25	-0.17	-0.42	0.17	0.04	0.76	-0.21	-0.23	-0.21	1.15	up	-1.23	-1.31	-2.39	0.83	0.15	-2.74	-1.50	-1.46	-3.39	0.009	ribosomal protein
sp|P18544|ARGD Acetylornithine aminot	-0.04				-0.06	-0.80	-0.36	-0.48	-0.35	1.27	up	-0.13				-0.12	-1.09	-2.46	-2.54	-3.36	0.010	arginine synthesis
sp|Q06405|ATPK ATP synthase subunit f		-0.60	-0.43	0.33	-0.30	-0.31	-0.19		-0.30	1.23	up		-2.41	-0.99	0.73	-1.90	-1.74	-1.01		-3.34	0.011	mitochondrial proton-transporting ATP synthase complex
sp|P53858|BNI4 Protein BNI4			-1.13	-0.53					-0.77	1.71	up			-3.03	-1.74					-3.26	0.014	cell morphogenesis regulation
sp|P36139|PET10 Protein PET10	-0.47		-0.55	-0.03	-0.24	0.42	-0.23	-0.38	-0.28	1.21	up	-1.96		-1.94	-0.10	-0.76	1.11	-1.57	-2.15	-3.23	0.015	mitochondrial membrane organization
sp|P32288|GLNA Glutamine synthetase O	0.02	-0.50	-0.04	-0.39	0.26	-0.76	-0.09	-0.17	-0.20	1.14	up	0.08	-3.11	-0.26	-2.12	1.44	-3.68	-0.61	-1.13	-3.23	0.015	glutamine biosynthesis
sp|Q04304|YMY0 UPF0659 protein YMR090	-0.70		-0.16	-0.11	-0.21	-0.16	-0.28		-0.25	1.19	up	-2.79		-0.88	-0.55	-1.22	-0.73	-1.92		-3.23	0.015	oxidative stress response, regulated by Yap1
sp|P49723|RIR4 Ribonucleoside-diphosp	-0.20	-0.14	-0.12	-0.14	-0.14	-0.35	-0.21	-0.20	-0.17	1.12	up	-0.98	-1.22	-1.05	-1.01	-0.64	-1.46	-1.49	-1.36	-3.18	0.018	deoxyribonucleoside diphosphate metabolic process
sp|P41805|RL10 60S ribosomal protein	-0.13	-0.04	-0.25	-0.01	0.35	-0.31	-0.11	-0.08	-0.16	1.12	up	-0.58	-0.26	2.35	-0.07	2.29	-1.76	-0.74	-0.55	-3.12	0.021	ribosomal protein
sp|P38113|ADH5 Alcohol dehydrogenase	-0.01	0.36	0.27	-0.88			-0.78		-0.35	1.28	up	-0.03	1.38	0.70	-1.88			-4.84		-3.10	0.023	ethanol biosynthetic process
sp|Q01217|ARG56 Protein ARG5,6, mitoc	-0.17	0.02					-0.47	-0.29	-0.28	1.21	up	-0.83	0.11					-3.32	-1.44	-3.08	0.024	arginine synthesis
sp|P22217|TRX1 Thioredoxin-1	0.21	0.04	-0.26	-0.24	-0.40	-0.31	0.06	0.07	-0.16	1.11	up	0.95	0.28	-2.33	-1.78	-3.14	-1.99	0.42	0.36	-3.08	0.024	oxidative stress response
sp|P40515|FIS1 Mitochondria fission 1	-0.27						-0.49		-0.43	1.35	up	-0.95						-2.97		-3.05	0.026	mitochondrial membrane organization
sp|P47122|NPA3 GTPase NPA3	-0.09						0.65		-0.48	1.39	up	-0.32						-3.43		-3.03	0.027	GTP binding
sp|Q07163|TY1AB Transposon TyH3 Gag-P	0.03	0.02	-0.08	-0.21	-0.22	-0.66	-0.15	-0.21	-0.15	1.11	up	0.17	0.18	-0.85	-1.79	-1.22	-3.21	-1.12	-1.43	-3.02	0.028	transposition
sp|P53221|RL26B 60S ribosomal protein	-0.15		-0.38	0.02			-0.36	-0.29	-0.25	1.19	up	-0.51		-1.67	0.09			-2.31	-1.66	-2.96	0.033	ribosomal protein
sp|P54861|DNM1 Dynamin-related protei	-0.23	-0.42					-0.29		-0.31	1.24	up	-1.04	2.03					-1.56		-2.93	0.036	mitochondrial membrane organization
sp|P48361|ASK10 Protein ASK10	-1.04						-0.36		-0.56	1.48	up	-2.93						-1.57		-2.91	0.039	oxidative stress response
sp|Q12000|TMA46 Translation machinery	-0.63						-0.34		-0.40	1.32	up	-1.98						-2.21		-2.85	0.045	ribosome biogenesis

Table 3. continued

Protein Description[a]	Corrected $\log_2$-ratio: 18O HR	18O LR	iTRAQ PQD 114-115	iTRAQ PQD 116-117	iTRAQ TOFTOF 114-115	iTRAQ TOFTOF 116-117	SILAC HR	SILAC LR	x'_q	Fold change		z_{eq}: 18O HR	18O LR	iTRAQ PQD 114-115	iTRAQ PQD 116-117	iTRAQ TOFTOF 114-115	iTRAQ TOFTOF 116-117	SILAC HR	SILAC LR	z'_q	FDR'_q	Comments
sp|P29952|MPI Mannose-6-phosphate iso	0.16	0.14	0.14	0.13	0.25	0.13	0.20	0.20	0.17	1.13	down	0.80	1.14	0.72	0.58	1.54	0.70	1.46	1.32	3.01	0.047	cell wall biosynthesis
sp|P25383|YC21A Transposon Ty2-C Gag							0.47	0.24	0.37	1.29	down							2.86	1.36	3.03	0.046	transposition
sp|P32775|GLGB 1,4-alpha-glucan-branc	0.30		0.55	-0.14			0.34	0.19	0.28	1.21	down	1.29		1.61	-0.42			2.40	1.10	3.04	0.044	glycogen metabolic process
sp|P40510|SER33 D-3-phosphoglycerate	0.45	0.32	0.17	-0.01			0.19		0.23	1.17	down	2.12	2.28	1.00	-0.06			1.39		3.08	0.040	serine family amino acid biosynthetic process
sp|P15019|TAL1 Transaldolase	0.02	0.01	0.04	0.21	0.15	0.24	0.24	0.23	0.13	1.10	down	0.12	0.09	0.50	1.88	1.35	1.73	1.74	1.54	3.08	0.041	pentose-phosphate shunt
sp|P11154|PYC1 Pyruvate carboxylase 1	0.35	0.20	0.00	0.07			0.25	0.41	0.17	1.13	down	1.68	1.55	0.00	0.56			1.82	2.71	3.09	0.040	gluconeogenesis
sp|P00925|ENO2 Enolase 2	-0.04	0.13	0.09	0.12	0.05	0.22	0.21	0.20	0.12	1.09	down	-0.22	1.16	0.90	0.95	0.66	2.38	1.53	1.36	3.11	0.039	gluconeogenesis
sp|P38934|BFR1 Nuclear segregation pr	0.07	0.31	0.08	0.21	0.22	0.14	0.15	0.19	0.16	1.11	down	0.36	1.38	0.84	1.72	1.51	0.81	1.06	1.25	3.11	0.040	mRNA metabolic process
sp|Q08224|THI20 Phosphomethylpyrimidi	0.43	0.14					0.38	0.36	0.32	1.25	down	1.10	0.69					2.47	1.90	3.20	0.030	mRNA metabolic process
sp|Q3E841|YN034 Uncharacterized prote	0.10	0.18	-0.26	0.40	0.39	0.01	0.48		0.23	1.17	down	0.40	1.17	-1.24	1.63	2.15	0.06	3.43		3.22	0.029	
sp|P33754|SEC66 Translocation protein							-0.26	1.16	0.44	1.36	down							-1.34	5.97	3.25	0.026	cell growth
sp|P37291|GLYC Serine hydroxymethyltr	0.15	0.15	0.09	0.12	0.15	0.16	0.32	0.20	0.16	1.11	down	0.72	1.20	0.99	1.07	0.91	0.85	2.33	1.35	3.28	0.025	serine family amino acid metabolic process
sp|P04807|HXK8 Hexokinase-2	0.01	0.13	0.07	0.21	0.41	0.38	0.24	0.20	0.18	1.13	down	0.08	1.09	0.50	1.32	2.00	1.58	1.78	1.35	3.29	0.024	hexose import
sp|P07149|FAS1 Fatty acid synthase su	0.08	0.08	0.10	0.16	0.08	0.21	0.21	0.19	0.13	1.10	down	0.43	0.73	1.24	1.50	0.89	1.79	1.52	1.29	3.29	0.025	fatty acid synthase activity
sp|P00560|PGK Phosphoglycerate kinase	-0.12	0.11	0.15	0.12	0.05	0.20	0.12	0.16	0.11	1.08	down	-0.62	0.99	1.93	1.13	0.82	2.52	0.86	1.13	3.32	0.023	gluconeogenesis
sp|P39003|HXT6 High-affinity hexose t		0.14	0.13	-0.11			0.74		0.30	1.23	down		0.62	0.74	-0.58			4.93		3.34	0.022	hexose transport
sp|P40043|YEP7 Uncharacterized protei	0.11	0.10	-0.22	0.12			0.53	0.65	0.28	1.21	down	0.39	0.60	-0.77	0.47			3.53	2.45	3.37	0.021	energy reserve metabolic process
sp|P11632|NHP6A Non-histone chromosom			0.64	0.21			0.50		0.43	1.34	down			1.78	0.77			3.01		3.46	0.015	polymerase III transcriptional preinitiation complex asser
sp|P07283|PMM Phosphomannomutase	0.07	0.22	0.20	0.24	0.00	-0.08	0.30	0.31	0.18	1.13	down	0.36	1.65	1.84	1.77	-0.02	-0.47	2.12	2.68	3.52	0.013	hexose biosynthetic process
sp|P46989|ATG27 Autophagy-related pro	-0.19				-0.01	0.89	0.72		0.46	1.38	down	-0.69				-0.03	1.90	4.28		3.54	0.012	vacuole organization
sp|P21524|RIR1 Ribonucleoside-diphosp	0.54	0.19	0.24	0.20			0.41		0.33	1.26	down	2.19	0.99	0.94	0.70			2.71		3.55	0.012	DNA replication
sp|Q03497|STE20 Serine/threonine-prot		0.70					0.37		0.51	1.43	down		3.24					1.92		3.59	0.011	positive regulation of protein kinase activity
sp|P19358|METK2 S-adenosylmethionine	0.26	0.15	0.15	0.18	0.19	0.09	0.31	0.26	0.20	1.15	down	1.22	1.02	1.08	1.15	1.16	0.43	2.23	1.75	3.62	0.010	sulfur compound biosynthetic process
sp|P46367|ALDH4 Potassium-activated a	0.05	0.00	0.03	0.06	0.25	0.40	0.44	0.38	0.18	1.13	down	0.25	0.00	0.32	0.44	1.85	2.48	3.17	2.54	3.68	0.008	alcohol metabolic process
sp|P47120|DOHH Deoxyhypusine hydroxyl	0.27		0.22	0.48	1.19	-0.64	0.17		0.29	1.22	down	1.17		1.58	2.99	2.92	-1.24	1.18		3.70	0.008	translation
sp|P04046|PUR1 Amidophosphoribosyltra	-0.01				1.29	-0.28	0.28	0.76	0.37	1.30	down	-0.06				2.26	-0.32	1.95	6.05	3.77	0.006	purine base metabolic process
sp|P54115|ALDH6 Magnesium-activated a	0.13	0.13	0.17	0.23	0.22	0.24	0.17	0.08	0.17	1.13	down	0.65	1.10	1.97	2.00	1.52	1.36	1.22	0.53	3.80	0.006	NADPH regeneration
sp|P01095|IPB2 Protease B inhibitors	0.24	-0.02	0.24	0.53	0.24	0.21	0.26	0.29	0.26	1.20	down	0.81	-0.08	1.51	2.74	0.98	0.66	1.85	1.76	3.83	0.006	vacuole organization
sp|Q06408|ARO10 Transaminated amino a	0.52						0.59		0.57	1.48	down	2.83						3.35		3.91	0.004	L-phenylalanine catabolic process
sp|P00812|ARGI Arginase	0.55		-0.07	0.32			0.77		0.51	1.42	down	3.78		-0.23	1.04			4.34		4.08	0.002	arginine catabolic process
sp|P39976|DLD3 D-lactate dehydrogenas	0.14	0.20	0.30	0.31	0.22	0.05	0.36	0.39	0.26	1.20	down	0.69	1.53	1.41	1.53	1.10	0.22	2.61	2.25	4.19	0.001	D-lactate dehydrogenase (cytochrome) activity
sp|Q04371|YMR7 UPF0364 protein YMR027	0.53		1.15	-0.17			0.04	0.57	0.39	1.31	down	1.97		0.24	-0.83			0.27	3.36	4.50	4.E-04	
sp|Q08911|FDH1 Formate dehydrogenase	0.24		0.40	0.52			0.55		0.47	1.39	down	0.99		1.67	2.01			3.07		4.42	2.E-04	NADH metabolic process
sp|P43616|CPG1 Glutamate carboxypepti	0.02	0.38	0.23	0.35	-0.21	0.89	0.31	0.36	0.30	1.23	down	0.10	2.82	1.53	2.16	-0.58	2.64	2.25	2.16	4.88	7.E-05	dipeptidase activity
sp|P00924|ENO1 Enolase 1	-0.04	0.13	0.17	0.27	0.13	0.25	0.30	0.32	0.19	1.14	down	-0.21	1.21	2.05	2.46	1.79	2.75	2.15	2.22	5.22	1.E-05	gluconeogenesis
sp|P07991|OAT Ornithine aminotransfer	0.07	0.10	0.60	-0.03	1.43	0.83	0.53	0.64	0.36	1.29	down	0.33	0.73	2.54	-0.12	1.83	0.74	3.94	4.15	5.25	1.E-05	arginine catabolic process
sp|P16474|GRP78 78 kDa glucose-regula	0.06	0.27	0.10	0.27	0.29	0.22	0.30	0.22	0.22	1.16	down	0.29	2.37	1.14	2.51	3.01	1.86	2.14	1.51	5.31	9.E-06	response to unfolded protein
sp|P38009|PUR92 Bifunctional purine b	0.56	0.15	0.53	0.21			0.50	0.46	0.38	1.30	down	2.68	1.05	2.21	0.89			3.61	2.68	5.36	8.E-06	purine base metabolic process
sp|P17967|PDI Protein disulfide -isome	0.08	0.16	0.29	0.25	0.23	0.73	0.26	0.25	0.26	1.20	down	0.42	1.32	2.94	2.07	1.40	3.86	1.86	1.67	5.49	4.E-06	endoplasmic reticulum lumen
sp|Q03529|SCS7 Inositolphosphorylcera		0.69	-0.14	0.36			0.88		0.69	1.62	down		3.16	-0.35	0.83			5.33		5.75	1.E-06	lipid biosynthesis
sp|P11986|INO1 Inositol-3-phosphate s	0.27	0.24	0.27	0.15	0.35	0.12			0.25	1.19	down	1.38	2.17	2.93	1.29	4.37	1.21			5.79	1.E-06	lipid biosynthesis
sp|P07262|DHE4 NADP-specific glutamat	0.11	0.16	0.31	0.28	0.13	0.13	0.56	0.55	0.25	1.19	down	0.59	1.35	3.18	2.34	1.62	1.22	4.07	3.71	6.19	1.E-07	synthesis of α-Ketoglutarate
sp|P40989|FKS2 1,3-beta-glucan syntha		1.55	-0.14				0.25		1.04	2.06	down		8.23	-0.30				0.80		6.89	1.E-09	cell wall biosynthesis
sp|P40825|SYAC Alanyl-tRNA synthetase	0.41	0.32	0.23	0.43	0.40	0.65	0.46	0.53	0.40	1.32	down	2.13	2.49	2.18	3.40	2.97	3.84	3.32	3.20	8.16	0.E+00	translation
sp|P10127|ADH4 Alcohol dehydrogenase	0.39	0.42	0.38	0.56	0.56	0.47	0.69	0.65	0.51	1.42	down	1.94	3.33	3.39	3.99	5.20	3.68	4.93	4.30	10.99	0.E+00	alcohol biosynthetic process

[a]Proteins changing in abundance at $FDR'_q < 0.05$ after data integration and detected in two or more experiments. The list does not include the protein sp|P47912|LCF4_YEAST Long-chain-fatty-acid--CoA ligase 4, which is significantly increased after data integration, but is also increased in the SILAC-HR and SILAC-LR null-hypothesis experiments. Proteins are sorted by z'_q. The magnitudes of the expression change and of the standardized variable are shaded according to the color scale at the top.

units of standardized deviation; using this variable, it was possible to analyze the joint behavior of all of the spectra in terms of a unique distribution (Figure 1, inset). When the quantitative data obtained from the analysis of the NH samples were processed using this procedure, the standardized variables describing the variability between different spectra within the same peptide (z_{eqps}, Figure 1A) closely followed a normal distribution with zero mean and unit variance (Figure 2, left column), demonstrating the accuracy of the model and the validity of the NH at the spectrum level for all SIL methods and MS machines.

Errors produced at the peptide and protein levels are treated in the WSPP model by assuming the simplest case (i.e., that they are normally distributed with zero mean and constant variance). This statement is reflected by the two fixed resistances in the peptide and protein branches (σ^2_{Pe} and σ^2_{Se}, respectively) in the electrical circuit (Figure 1). The validity of this double assumption was demonstrated by the accuracy with which the corresponding standardized variables (z_{eqp} and z_{eq}) follow the expected normal distribution with zero mean and unit variance in all SIL methods (Figure 2, center and right columns). In the two SILAC cases, because of the metabolic nature of this kind of labeling, it was not possible to compare identical replicates, and a pseudonull hypothesis (A* vs A) was used instead. Although, as expected, some minor abundance changes were detected (Table 1), the distribution of standardized variables at the peptide and protein levels from the SILAC experiments were also in excellent agreement with the predictions of the model (Figure 2, center and right columns).

Interestingly, the variances estimated at the three levels in the different SIL approaches (Tables 1 and 2) reflect the specific characteristics of each method. For instance, iTRAQ quantifications by PQD have a greater variance at the spectrum level than those obtained by TOF/TOF, consistent with the fact that the latter produce higher-intensity reporter ions that produce more accurate readings. However, their peptide variances are similar, as expected for methods that share a common peptide-preparation procedure. Likewise, SILAC variances at the peptide level are practically zero, as expected for a predigestion labeling method, whereas the other methods share a similar nonzero variance at the protein level. Finally, SILAC variances at the spectrum level were also the lowest among all SIL methods, reflecting the facts that SILAC quantification is directly performed in the MS spectrum and that the mass difference between labeled and unlabeled species is higher than that produced by ^{18}O labeling, thus minimizing quantification errors resulting from isotopomer superposition.

To test the WSPP algorithm further, we also analyzed whether it was possible to construct a statistical model containing fewer parameters. As described in Supporting Information section 3.1, the model had to take into account three sources of variance to be able to describe the results of all of the NH experiments, producing a negligible number of false protein abundance changes in all cases (Supporting Informa-

tion Figure 3). Finally, the performance of the WSPP model was compared with that of other commonly used statistical models specifically developed for ^{18}O[40], iTRAQ,[25,30] or SILAC[33] (Supporting Information section 3.2). We analyzed the accuracy of each model to describe the results obtained in the NH experiments (Supporting Information Figures 5−7), compared the performance of the WSPP model with that of a weighted least-squares method for iTRAQ[30] in terms of accuracy and FDR using a yeast background with spiked-in proteins at different concentrations and ratios (Supporting Information Figure 8), and performed a side-by-side comparison of WSPP and MaxQuant to determine protein-abundance changes produced by H_2O_2 treatment (Supporting Information Tables 5 and 6). Our results indicate that in spite of their general applicability the performance of the WSPP model was at least similar to that of these previous models.

WSPP Model Allows Comparison and Coherent Integration of Results Obtained from All Quantification Approaches, Increasing the Statistical Power of Protein Quantification

The analysis using the WSPP model of the relative quantification of control versus H_2O_2-treated samples produced a set of statistical parameters very similar to the NH experiments (compare Table 1 with Table 2), highlighting the robustness of the algorithm in a real quantitative experiment. The distribution of standardized variables at the spectral and peptide levels was also in good agreement with the null hypothesis (Supporting Information Figure 4), reinforcing the validity of the model. At the protein level, a significant number of quantifications (ranging from 5 to 39) behaved as outliers at a 5% FDR (Figure 3A−F and Tables 1 and 2), reflecting protein-abundance changes produced by the H_2O_2 treatment. Without an appropriate statistical hypothesis, these results could be only compared and validated for the 16 proteins that changed their abundance in at least two experiments. As shown in Supporting Information Table 3, there was a good agreement between the different experiments. These results were further confirmed by an independent analysis of some of the changing proteins by using label-free quantification (Supporting Information Table 4 and Supporting Information Figure 9). In spite of the consistency of these results, the observed changes provided very limited biological information, indicating only the activation of a first-line defense against oxidative conditions, probably through the H_2O_2-responsive transcriptional activators Yap1p, Skn7p, and Msn2/4p42, which induced expression of thiol homeostasis proteins such as cytochrome c peroxidase, thioredoxin 2, and glutathione peroxidase 2.

To judge the reproducibility and dispersion of results among different experiments and to integrate the quantitative information from all of them, the WSPP model introduces a fourth layer of analysis. In this layer, protein quantifications are averaged from the different experiments according to error propagation theory (Supporting Information), a procedure mathematically equivalent to setting up protein batteries in parallel (Figure 1B). Demonstrating the validity of this fourth layer, the distribution of the standardized variables describing the variability between different NH experiments within the same protein (Figure 3G) and between different proteins (Figure 3H) were exactly as expected from theory, and no significant protein-abundance changes were detected at a 5% FDR (Figure 3I). The interexperiment distribution of protein data from the control versus H_2O_2-treatment experiments also followed the expected trend (Figure 3K), demonstrating that quantifications from all of the experiments were, again, reproducible and therefore can be integrated (Table 3). The protein distribution, however, clearly deviated from the NH trend (Figure 3L), and 123 significant protein-abundance changes at 5% FDR were revealed in the tornado plot (Figure 3M). Hence, the number of significant abundance changes increased almost 1 order of magnitude when the results from the different experiments were integrated in the framework of the WSPP model, whereas there were no significant abundance changes when results from the NH experiments were integrated using the same approach. The consistency in the quantitative behavior of these significantly changing proteins in the original experiments is illustrated in Supporting Information Figure 10. This increase in statistical significance afforded by data integration allowed not only the detection of more upregulated proteins implicated in primary stress response but also detection of numerous proteins related to the ribosome and the mitochondrial membrane as well as downregulated proteins implicated mostly in metabolic functions such as glycolysis and gluconeogenesis (Table 3), a finding that agrees with other works.[41−44]

■ DISCUSSION

In this study, we present WSPP, a general statistical framework for the analysis of quantitative proteomics results, and we demonstrate that the statistical model very accurately describes the technical variability of data for a representative set that includes the most common SIL methods. In addition, the model allows a systematic comparison and integration of data from different experiments. The general validity of the model was demonstrated by the analysis of 16 quantitative experiments performed using six different combinations of SIL and MS approaches (Tables 1 and 2). The distribution of the data were analyzed at three levels (spectrum, peptide, and protein) so that a total of 48 experimental distributions were confronted against 18 different null hypotheses. This statistical framework efficiently resolves the problems of variance heterogeneity, data integration, and under-sampling and provides a statistically sound method for testing the quality of quantitative experiments and detecting experimental deviations. Moreover, our results reinforce the idea that quantitative SIL experiments, if properly performed and analyzed, are highly reproducible irrespective of the labeling technique or MS platform used.

At the spectral level, the model analyzes variance using a strategy different from variance-stabilization normalization procedures used to treat microarray[32] and iTRAQ data.[25,31] Instead of transforming the data, WSPP uses a two-parameter function to model the behavior of variance, and from this function, it directly assigns a variance to each one of the quantifications, keeping the original readings. This two-parameter modeling of variance is similar to that followed in a previous work to treat iTRAQ data,[28] although in that work, the final analysis was made at the spectral level and no integration of the data at the peptide or protein levels was performed. The general applicability of our model to any SIL and MS combination confirms in the most general case a fundamental property of MS-based peptide-centric quantifications: the error produced during the quantitative SIL analysis of peptides generating the same intensity at a given MS detector, irrespective of their sequence or molecular structure, is constant and normally distributed. We believe that this property is of paramount importance in the field and simplifies interpretation

of quantitative data produced by MS. Finally, a model that analyzes specifically the variance at the spectral level has the advantage that it allows for separate control of the error produced during MS analysis and quantification from that produced at the time of peptide or protein preparation and thus may be used to check the proper functioning and calibration of the MS machines and even to detect chromatographic shifts resulting from incomplete coelution of peptide pairs. For instance, some of the MS conditions used in this work, such as the collision energy in PQD fragmentation, were optimized by selecting the ones that produced the minimum variance at the spectral level.

Our results also suggest that, at least using the protocol followed here, the error produced during peptide preparation can be considered constant and normally distributed, reinforcing previous results we obtained using other biological systems[15] and peptide-preparation methods.[34] In the general case, the null hypothesis formulated here provides the basis for testing the validity of this assumption for any peptide preparation method. Analysis of variance at the peptide level may be very useful in practice to detect deviations from the expected behavior during peptide preparation, such as those produced by artifacts like partial digestion or methionine oxidation.[15] However, it may also be used to assign a statistical significance level to abundance changes in postranslationally modified peptides, such as Cys sites subjected to oxidative modification, as we demonstrated in a recent work.[45] It could also be potentially used to detect effects because of the presence of SNPs or differential splicing. In addition, the protein layer can also be ignored in our model so that statistical analysis of abundance changes is directly performed at the peptide level, without grouping peptides into proteins.

Finally, the error at the protein level has been shown to follow the same trend in this and other biological systems,[15] making it a very convenient starting point from which to analyze other error sources, such as biological variance, which is highly dependent on experimental design and must be modeled separately in each case. Besides, because the variances at the protein level estimated in all of the SIL experiments performed in this work (Tables 1 and 2) are very similar to those calculated in previous studies using ^{18}O labeling in several biological models,[15,34] including tissue extracts, and in general are below 0.01, this value may be used as a reference to determine whether the protein variability in a given preparation is higher than that expected for a conventional protein-manipulation method. Thus, an increase in protein variance above the reference value not accompanied by a concomitant increase in peptide and spectrum variances indicates an increased heterogeneity in protein composition that is not related to peptide manipulation or MS quantification. This heterogeneity may indicate technical problems related, for instance, to a protein-preparation protocol involving too many steps, but it may also reflect a high biological variability between the samples that are compared, as we have found in a previous work[46] and also when comparing human samples extracted from different individuals (unpublished data). Note that the majority of existing models to analyze SIL data integrate the data intro protein averages, without taking into account the variance at the lower levels, and then analyze the distribution of protein quantifications as a whole;[25,26,31] in doing so, all of the technical error sources are comprised in only one random variable, and it is not possible to interpret results in terms of the variance at the protein level.

One of the most important characteristics of the WSPP model is that it provides a general framework to make a full integration of quantitative and error information from one level to a superior level. When several spectra are integrated into a peptide average, the model takes into account the variance associated with each one of the spectra according to error propagation theory so that the most accurate have a more significant contribution. This procedure simplifies the interpretation of data because quantifications of poor quality have a negligible effect on the peptide average and do not need to be eliminated from the analysis. However, the variance assigned to each one of the peptides takes into account not only the variance at the spectrum level, which is diminished because of averaging several spectra, but also the intrinsic variance associated with the process of peptide preparation. The same is done when peptides are integrated into proteins; although peptide averaging diminishes the variance carried out from the spectrum and peptide levels, the protein average includes the variance produced by protein manipulation. In this sense, the constant variances at the spectrum, peptide, and protein levels can be conceived as asymptotic errors because they reflect the lowest error that can be achieved at each one of the levels (see horizontal bars in Figure 2). This property of data integration is illustrated in the equivalent circuit (Figure 1A), where resistances set in series reflect the additive nature of the variances associated to independent events, whereas those set in parallel reflect the effect of averaging on variances. Most importantly, this concept of data integration is of general validity and can be extended to upper levels, where the effect of averaging is reflected when protein readings from different experiments are considered together. The decrease in resistance (i.e., variance) of the data integrated at upper levels increase the statistical power to detect deviations from the null hypothesis. This explains why only a dozen of the significant protein-abundance changes are detected when the experiments were considered separately, whereas the alteration of more than 100 proteins becomes evident when the data are integrated at upper levels.

The use of weighted averages to calculate protein ratios is not new and reflects the current view that not all quantitative measurements have the same accuracy.[25−29,33] However, the weighting scheme followed by our model is different from other approaches in several aspects. Although other methods have been proposed that separately estimate the biological and technical variance components,[4,6] to the best of our knowledge and with the exception of its predecessor,[34] no models have been formulated previously for quantitative proteomics data that decompose technical variance into two or more components. Besides, in our model, the averages are calculated following error propagation theory so that the statistical weight with which each value contributes to the average is exactly the inverse of its local variance, and the variances of each one of the averaged values are known with accuracy. A similar kind of weighting by the error made at the spectrum and peptide levels was proposed in one of the earliest models put forth to analyze quantitative data produced by using stable isotope-coded affinity tags (ICAT);[47] however, the approach that was followed in that work only propagates the error produced at the time of quantification from the MS spectrum and does not take into account the variance introduced by further errors produced by peptide generation and protein manipulation. Finally, to the best of our knowledge, our weighting method is

the first one that has been demonstrated to be of general validity for a wide range of different SIL and MS approaches.

The WSPP model also resolves the problem of undersampling and at the same time provides a framework to integrate data and a robust algorithm to estimate variances, which are of general applicability. This is accomplished by using standardized variables at each one of the integration levels. These variables express $\log_2$ ratios in units of standard deviation, introducing a bias correction for the number of degrees of freedom, which depends on the number of elements that are used to compute the average (i.e, the number of peptides that are used to estimate a protein value). Using the standardized variables, all of the elements at a given level of integration can be analyzed together in a unique distribution that, under the null hypothesis, is expected to be a normal distribution with unit variance (Figure 1, central inset). We take advantage of this general property to make robust estimates of variances by an iterative method that is of general applicability for all integration levels and quantification approaches. We should note here that our approach to estimate variance is similar to that used by other authors,[28,31,33] although in these works, only the total technical variance was estimated. Analysis of the standardized variable is also very useful to detect the presence of outliers at any integration level; in the WSPP model, this may be done at seven different levels (z parameters in Figure 1), and in each one, it provides specific information. Thus, at the spectrum and peptide levels, outliers are indicative of incorrect quantifications produced by a variety of causes (for instance, peak coelution, bad fitting, or methionine oxidation[15,34]), as commented above. At the protein level, outliers indicate statistically significant abundance changes. However, at other levels of integration, an outlier may also indicate that an experimental replicate gives quantitative results that deviate from the other replicates more than expected by chance alone and the absence of outliers indicate that all the replicates behave as expected by the null hypothesis (as observed in this work). We should note here that this global conception of variances, which considers together the whole wealth of data, is in contrast with other approaches commonly followed to detect outliers and artifacts, like Dixon's test,[47] which locally analyze the distribution of the elements used to calculate a particular average and which have a very limited utility when the number of elements is very low (i.e., when a protein is quantified by only two peptides). In the WSPP model, all of the elements, even single hits, are assigned a local variance, and this is done on the basis of only four parameters that are estimated from the analysis of the whole collection of data.

■ CONCLUSIONS

We present a statistical framework that explains the behavior of data obtained using the most common SIL-MS approaches. The WSPP model allows a systematic comparison and integration of data from different SIL experiments, and, in spite of its general applicability, its performance is at least similar to other commonly employed methods. We demonstrate the importance of performing rigorous data integration to uncover subtle but widespread protein changes taking place in a proteome, using *S. cerevisiae* exposed to a low dose of H_2O_2 as an example. Thus, comparison of the results at the protein level revealed activation only of first-line responses in response to the treatment, whereas data integration from different SIL experiments using the WSPP framework uncovered more subtle changes in groups of proteins related to the ribosome and mitochondria and also to metabolic pathways. These results highlight the importance of establishing an adequate and validated statistical framework for the analysis of high-throughput quantitative data. We believe that the WSPP model may contribute to developing a general standard for the analysis of quantitative proteomics data obtained by stable isotope labeling.

■ ASSOCIATED CONTENT

Supporting Information

Additional materials and methods including details of the WSPP statistical model; evidence that three independent sources of variance are necessary in the WSPP model; comparison of the WSPP model with other existing algorithms; elements of the WSPP statistical model; definition of fitting weights; comparison of significant protein abundance changes in the control versus H_2O_2-treatment experiments; list of proteins and peptides subjected to a label-free parallel reaction monitoring; side-by-side comparison of protein abundance changes in the control versus H_2O_2-treatment SILAC-HR experiment determined by MaxQuant and WSPP; comparison of protein abundance changes obtained in the control versus H_2O_2-treatment experiments using WSPP and in the SILAC-HR experiment using MaxQuant; analysis of local normality; estimation of the weight constant for the different SIL experiments; evidence that three separate sources of variance at the spectrum, peptide, and protein level are needed to give a correct description of the different NH experiments; analysis at the spectrum, peptide, and protein levels of the controls versus H_2O_2-treatment experiments; analysis of the high- and low-resolution 18O NH experiments using UNiquant algorithm; analysis of the TOFTOF and PQD iTRAQ NH experiments using the variance-stabilizing normalization strategy; analysis of the SILAC-HR, pseudo-NH experiment using MaxQuant; performance of the WSSP model in the analysis of spike-in mixtures of known protein concentrations in a complex background; verification of some protein abundance changes in the control versus treatment experiments by label-free parallel reaction monitoring; comparison of changing ribosomal, oxidative stress, and metabolic proteins in the control versus treatment experiments; complete list of proteins quantified by the different SIL-MS approaches for the null hypothesis; complete list of proteins quantified by the different SIL-MS approaches for the controls versus treatment experiments; complete list of protein quantifications integrated from the different SIL-MS approaches for the null hypothesis and for the controls versus treatment experiments; and information about the meaning of the parameters used in QuiXoT. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jvazquez@cnic.es. Phone: (+34) 91 4531200. Fax: (+34) 91 4531245.

Author Contributions

◆These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants BIO2009-07990, BIO2009-11735, BFU2009-08004, SAF 2009-07520, and BIO2012-37926 from the Spanish Ministry of Science and Education, CAM BIO/0194/2006 (Cardiovrep) from the Madrid regional government, and an institutional grant from the Fundación Ramón Areces to the CBMSO. Grants RD06/0014/0030, RD12/0042/0021, RD06/0014/0005, and RD12/0042/0022 from the Red Temática de Investigación Cooperativa en Enfermedades Cardiovasculares (RECAVA/RIC, Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III, Ministry of Health) supported the research of J.V. and J.M.R. P.M.-A. is recipient of a fellowship from the Madrid regional government supported by the European Social Fund. We thank S. Bartlett for language editing.

■ REFERENCES

(1) Aebersold, R. A stress test for mass spectrometry-based proteomics. *Nat. Methods* **2009**, *6*, 411−412.
(2) Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R., 3rd; Bairoch, A.; Bergeron, J. J. Mass spectrometry in high-throughput proteomics: Ready for the big time. *Nat. Methods* **2010**, *7*, 681−685.
(3) Domon, B.; Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* **2010**, *28*, 710−721.
(4) Clough, T.; Key, M.; Ott, I.; Ragg, S.; Schadow, G.; Vitek, O. Protein quantification in label-free LC-MS experiments. *J. Proteome Res.* **2009**, *8*, 5275−5284.
(5) Chang, C. Y.; Picotti, P.; Huttenhain, R.; Heinzelmann-Schwarz, V.; Jovanovic, M.; Aebersold, R.; Vitek, O. Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol. Cell. Proteomics* **2012**, *11*, M111.014662-1−M111.014662-13.
(6) Daly, D. S.; Anderson, K. K.; Panisko, E. A.; Purvine, S. O.; Fang, R.; Monroe, M. E.; Baker, S. E. Mixed-effects statistical model for comparative LC-MS proteomics studies. *J. Proteome Res.* **2008**, *7*, 1209−1217.
(7) Karpievitch, Y.; Stanley, J.; Taverner, T.; Huang, J.; Adkins, J. N.; Ansong, C.; Heffron, F.; Metz, T. O.; Qian, W. J.; Yoon, H.; Smith, R. D.; Dabney, A. R. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **2009**, *25*, 2028−2034.
(8) Oberg, A. L.; Vitek, O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.* **2009**, *8*, 2144−2156.
(9) Polpitiya, A. D.; Qian, W. J.; Jaitly, N.; Petyuk, V. A.; Adkins, J. N.; Camp, D. G., 2nd; Anderson, G. A.; Smith, R. D. DAnTE: A statistical tool for quantitative analysis of -omics data. *Bioinformatics* **2008**, *24*, 1556−1558.
(10) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **1995**, *1*, 289−300.
(11) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **2002**, *1*, 376−386.
(12) Ong, S. E.; Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **2006**, *1*, 2650−2660.
(13) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154−1169.
(14) Mirgorodskaya, O. A.; Kozmin, Y. P.; Titov, M. I.; Korner, R.; Sonksen, C. P.; Roepstorff, P. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1226−1232.
(15) Bonzon-Kulichenko, E.; Perez-Hernandez, D.; Nunez, E.; Martinez-Acedo, P.; Navarro, P.; Trevisan-Herraz, M.; Ramos Mdel, C.; Sierra, S.; Martinez-Martinez, S.; Ruiz-Meana, M.; Miro-Casas, E.; Garcia-Dorado, D.; Redondo, J. M.; Burgos, J. S.; Vazquez, J. A robust method for quantitative high-throughput analysis of proteomes by ^{18}O labeling. *Mol. Cell. Proteomics* **2011**, *10*, M110.003335-1−M110.003335-14.
(16) Cox, J.; Mann, M. Is proteomics the new genomics? *Cell* **2007**, *130*, 395−398.
(17) Gan, C. S.; Chong, P. K.; Pham, T. K.; Wright, P. C. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). *J. Proteome Res.* **2007**, *6*, 821−827.
(18) Unwin, R. D.; Pierce, A.; Watson, R. B.; Sternberg, D. W.; Whetton, A. D. Quantitative proteomic analysis using isobaric protein tags enables rapid comparison of changes in transcript and protein levels in transformed cells. *Mol. Cell. Proteomics* **2005**, *4*, 924−935.
(19) Rodriguez-Suarez, E.; Gubb, E.; Alzueta, I. F.; Falcon-Perez, J. M.; Amorim, A.; Elortza, F.; Matthiesen, R. Virtual expert mass spectrometrist: iTRAQ tool for database-dependent search, quantitation and result storage. *Proteomics* **2010**, *10*, 1545−1556.
(20) Boehm, A. M.; Putz, S.; Altenhofer, D.; Sickmann, A.; Falk, M. Precise protein quantification based on peptide quantification using iTRAQ. *BMC Bioinf.* **2007**, *8*, 214-1−214-18.
(21) Hill, E. G.; Schwacke, J. H.; Comte-Walters, S.; Slate, E. H.; Oberg, A. L.; Eckel-Passow, J. E.; Therneau, T. M.; Schey, K. L. A statistical model for iTRAQ data analysis. *J. Proteome Res.* **2008**, *7*, 3091−3101.
(22) Oberg, A. L.; Mahoney, D. W.; Eckel-Passow, J. E.; Malone, C. J.; Wolfinger, R. D.; Hill, E. G.; Cooper, L. T.; Onuma, O. K.; Spiro, C.; Therneau, T. M.; Bergen, I.; Robert, H. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J. Proteome Res.* **2008**, *7*, 225−233.
(23) Oberg, A. L.; Mahoney, D. W. Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. *BMC Bioinf.* **2012**, *13*, S7-1−S7-18.
(24) Herbrich, S. M.; Cole, R. N.; West, J.; Keith, P.; Schulze, K.; Yager, J. D.; Groopman, J. D.; Christian, P.; Wu, L.; O'Meally, R. N.; May, D. H.; McIntosh, M. W.; Ruczinski, I. Statistical inference from multiple iTRAQ experiments without using common reference standards. *J. Proteome Res.* **2013**, *12*, 594−604.
(25) Karp, N. A.; Huber, W.; Sadowski, P. G.; Charles, P. D.; Hester, S. V.; Lilley, K. S. Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell. Proteomics* **2010**, *9*, 1885−1897.
(26) Lin, W. T.; Hung, W. N.; Yian, Y. H.; Wu, K. P.; Han, C. L.; Chen, Y. R.; Chen, Y. J.; Sung, T. Y.; Hsu, W. L. Multi-Q: A fully automated tool for multiplexed protein quantitation. *J. Proteome Res.* **2006**, *5*, 2328−2338.
(27) Shadforth, I. P.; Dunkley, T. P.; Lilley, K. S.; Bessant, C. i-Tracker: For quantitative proteomics using iTRAQ. *BMC Genomics* **2005**, *6*, 145-1−145-6.
(28) Zhang, Y.; Askenazi, M.; Jiang, J.; Luckey, C. J.; Griffin, J. D.; Marto, J. A. A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. *Mol. Cell. Proteomics* **2010**, *9*, 780−790.
(29) Bantscheff, M.; Boesche, M.; Eberhard, D.; Matthieson, T.; Sweetman, G.; Kuster, B. Robust and sensitive iTRAQ quantification on an LTQ orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2008**, *7*, 1702−1713.
(30) Mahoney, D. W.; Therneau, T. M.; Heppelmann, C. J.; Higgins, L.; Benson, L. M.; Zenka, R. M.; Jagtap, P.; Nelsestuen, G. L.; Bergen, I.; Robert, H.; Oberg, A. L. Relative quantification: Characterization of bias, variability and fold changes in mass spectrometry data from iTRAQ-labeled peptides. *J. Proteome Res.* **2011**, *10*, 4325−4333.
(31) Arntzen, M. O.; Koehler, C. J.; Barsnes, H.; Berven, F. S.; Treumann, A.; Thiede, B. IsobariQ: Software for isobaric quantitative proteomics using IPTL, iTRAQ, and TMT. *J. Proteome Res.* **2011**, *10*, 913−920.

(32) Huber, W.; von Heydebreck, A.; Sultmann, H.; Poustka, A.; Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **2002**, *18*, S96−S104.
(33) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.
(34) Jorge, I.; Navarro, P.; Martinez-Acedo, P.; Nunez, E.; Serrano, H.; Alfranca, A.; Redondo, J. M.; Vazquez, J. Statistical model to analyze quantitative proteomics data obtained by $^{18}O/^{16}O$ labeling and linear ion trap mass spectrometry: Application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells. *Mol. Cell. Proteomics* **2009**, *8*, 1130−1149.
(35) Gruhler, A.; Olsen, J. V.; Mohammed, S.; Mortensen, P.; Faergeman, N. J.; Mann, M.; Jensen, O. N. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* **2005**, *4*, 310−327.
(36) Martinez-Bartolome, S.; Navarro, P.; Martin-Maroto, F.; Lopez-Ferrer, D.; Ramos-Fernandez, A.; Villar, M.; Garcia-Ruiz, J. P.; Vazquez, J. Properties of average score distributions of SEQUEST: The probability ratio method. *Mol. Cell. Proteomics* **2008**, *7*, 1135−1145.
(37) Vizcaino, J. A.; Cote, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; O'Kelly, G.; Schoenegger, A.; Ovelleiro, D.; Perez-Riverol, Y.; Reisinger, F.; Rios, D.; Wang, R.; Hermjakob, H. The PRoteomics IDEntifications (PRIDE) database and associated tools: Status in 2013. *Nucleic Acids Res.* **2013**, *41*, D1063−D1069.
(38) Lopez-Ferrer, D.; Ramos-Fernandez, A.; Martinez-Bartolome, S.; Garcia-Ruiz, P.; Vazquez, J. Quantitative proteomics using $^{16}O/^{18}O$ labeling and linear ion trap mass spectrometry. *Proteomics* **2006**, *6*, S4−S11.
(39) Ramos-Fernandez, A.; Lopez-Ferrer, D.; Vazquez, J. Improved method for differential expression proteomics using trypsin-catalyzed ^{18}O labeling with a correction for labeling efficiency. *Mol. Cell. Proteomics* **2007**, *6*, 1274−1286.
(40) Huang, X.; Tolmachev, A. V.; Shen, Y.; Liu, M.; Huang, L.; Zhang, Z.; Anderson, G. A.; Smith, R. D.; Chan, W. C.; Hinrichs, S. H.; Fu, K.; Ding, S. J. UNiquant, a program for quantitative proteomics analysis using stable isotope labeling. *J. Proteome Res.* **2011**, *10*, 1228−1237.
(41) Herrero, E.; Ros, J.; Belli, G.; Cabiscol, E. Redox control and oxidative stress in yeast cells. *Biochim. Biophys. Acta* **2008**, *1780*, 1217−1235.
(42) Godon, C.; Lagniel, G.; Lee, J.; Buhler, J. M.; Kieffer, S.; Perrot, M.; Boucherie, H.; Toledano, M. B.; Labarre, J. The H_2O_2 stimulon in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **1998**, *273*, 22480−22489.
(43) Molina-Navarro, M. M.; Castells-Roca, L.; Belli, G.; Garcia-Martinez, J.; Marin-Navarro, J.; Moreno, J.; Perez-Ortin, J. E.; Herrero, E. Comprehensive transcriptional analysis of the oxidative response in yeast. *J. Biol. Chem.* **2008**, *283*, 17908−17918.
(44) McDonagh, B.; Ogueta, S.; Lasarte, G.; Padilla, C. A.; Barcena, J. A. Shotgun redox proteomics identifies specifically modified cysteines in key metabolic enzymes under oxidative stress in *Saccharomyces cerevisiae*. *J. Proteomics* **2009**, *72*, 677−689.
(45) Martinez-Acedo, P.; Nunez, E.; Gomez, F. J.; Moreno, M.; Ramos, E.; Izquierdo-Alvarez, A.; Miro-Casas, E.; Mesa, R.; Rodriguez, P.; Martinez-Ruiz, A.; Dorado, D. G.; Lamas, S.; Vazquez, J. A novel strategy for global analysis of the dynamic thiol redox proteome. *Mol. Cell. Proteomics* **2012**, *11*, 800−813.
(46) Bonzon-Kulichenko, E.; Martinez-Martinez, S.; Trevisan-Herraz, M.; Navarro, P.; Redondo, J. M.; Vazquez, J. Quantitative in-depth analysis of the dynamic secretome of activated Jurkat T-cells. *J. Proteomics* **2011**, *75*, 561−571.
(47) Li, X. J.; Zhang, H.; Ranish, J. A.; Aebersold, R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 6648−6657.

3. Third article

3.1 (English) A novel systems-biology algorithm for the analysis of coordinated protein responses using quantitative proteomics

Most systems biology models are based on genomics/transcriptomics data. Proteomics-based systems biology models have been explored but the potential given by the wealth of data, especially in current high-throughput quantitative proteomics, has been largely undeveloped. In this paper we present a further development of the WSPP statistical model introduced in the previous paper, called *Generic Integration Algorithm* (GIA), using it to establish the basic structure of the *Systems Biology Triangle* (SBT). To our knowledge, the SBT is the first algorithm used to detect protein coordinated behaviour in pairwise quantitative proteomics experiments. In other words, it tells us which categories (obtained from published protein databases) contain proteins whose relative abundance is changing as a group with other proteins equally classified, and, at the same time, detecting which proteins are changing differently, meaning they are outliers of the category to which they belong. The concept of protein coordination is defined and studied in detail to achieve a better understanding of the global behaviour of biological systems from the proteomic point of view. Additionally, a measurable dimensionless quantity, the *degree of coordination*, is presented in order to analyse such experiments, expressing the ratio of proteins changing together with regard to those that are acting as outliers in their groups.

The SBT model was tested by analysing a total of eight different biological models subjected to quantitative proteomics by several stable isotope labelling methods. These included a) yeast cells treated with H_2O_2, b) untreated yeast cells (Navarro, 2014); c) human bone marrow mesensphere secretome obtained from cells cultured in the presence of human serum or chicken embryo extract (Isern, 2013); d) liver mitochondrial-associated endoplasmic reticulum membranes (MAM) from Cav-1 knockout or wild type mice (Wieckowski, 2009); e) total fibroblast lysates from Zmpste24 KO or wild-type mice (Varela, 2005); f) total heart extracts from pigs after 1 hour post-infarction reperfusion (Danielsen, 2007; Garcia-Prieto, 2014); g) mouse lymphocytes obtained from HDAC6 KO and wild-type mouse spleen (Gonzalez-Granado, 2014) and h) primary mouse vascular smooth muscle cells (VSMC) treated with angiotensin-II (AngII); in VSMC samples six independent biological preparations

were used. Details about the null hypothesis used as a reference to determine significance, randomisation and statistical tests used are described in the paper.

The paper presented has been prepared in tight collaboration with Fernando García-Marqués, who is also a first author. My contribution is presented in the first part, and consists of i) the modification of the WSPP statistical model, ii) the Generic Integration Algorithm (GIA), iii) the software that supports it (SanXoT), iv) the Systems Biology Triangle (SBT) and v) the bioinformatic analysis of the eight models specified in the previous paragraph. The second part of the paper consists of a very relevant application of my work, in which a time-course high-throughput quantitative proteomics experiment using iTRAQ 8-plex is carried out to analyse the short term effect of angiotensin-II (AngII) to the vascular smooth muscle cells of the abovementioned sample. The algorithm is particularly successful to detect functional protein alterations produced by the coordinated action of the proteome, as well as in finding outlier proteins compared to other proteins equally classified, showing relevant information for the biological interpretation. The application of the SBT algorithm to the quantitative data revealed that AngII produced on VSMC an array of increasingly coordinated protein responses which were reproducibly observed along time. These results are of remarkable interest in the identification of therapeutic targets for the treatment of cardiovascular diseases, and more specifically, they have important mechanistic implications in understanding the role of AngII in vascular remodelling. Further biological interpretation of these results, which is available in the second part of the paper, is not presented in this introduction, as it is part of the PhD Thesis of Fernando García-Marqués.

In conclusion, the algorithm presented is unique to uncover the global picture of the changes taking place in a proteome, especially revealing the regulation of signalling and metabolic pathways observable in pairwise quantitative proteomics experiments, greatly improving the results provided by other current approaches, such as enrichment network analysis and functional class scoring methods.

3.2 (Español) Un algoritmo de biología de sistemas innovador para analizar la respuesta coordinada de las proteínas mediante proteómica cuantitativa

La mayoría de modelos de biología de sistemas se basan en datos de genómica/transcriptómica. Se han explorado otros modelos en los que se trata la biología de sistemas desde un enfoque proteómico, pero el potencial de la gran cantidad de datos disponibles, especialmente aquellos relacionados con proteómica cuantitativa de alto rendimiento, sigue desaprovechándose. En este artículo presentamos un desarrollo adicional del modelo estadístico WSPP mencionado en el artículo anterior, llamado *Algoritmo de Integración Genérico* (GIA), y lo utilizamos para establecer la estructura del *Triángulo de la Biología de Sistemas* (SBT). El SBT es, según nuestro conocimiento, el primer algoritmo que existe para detectar el comportamiento coordinado de proteínas en experimentos de proteómica cuantitativa. En otras palabras, nos indica qué categorías (encontradas en bases de datos proteicas publicadas) incluyen proteínas cuya abundancia relativa cambia en grupo junto con otras proteínas clasificadas de igual forma, detectando a la vez aquellas proteínas que varían de manera distinta, es decir, que son atípicas con respecto a la categoría a la que pertenecen. Se define el concepto de coordinación para proteínas, y se estudia en detalle para alcanzar una comprensión mayor del comportamiento global de biología de sistemas desde un enfoque proteómico. Adicionalmente, se presenta una cantidad mensurable, el *grado de coordinación*, con el objetivo de analizar dichos experimentos, que expresa la relación entre las proteínas que cambian juntas con respecto a aquellas cuyas pautas no se corresponden a las del grupo al que pertenecen.

Para testar el modelo SBT, se han analizado un total de ocho modelos biológicos diferentes, con los que se ha hecho proteómica cuantitativa utilizando varios métodos de marcaje isotópico estable. Estos incluyeron a) células de levadura tratadas con H_2O_2, b) células de levadura sin tratar (Navarro, 2014); c) secretoma de mesenesferas de tuétano de hueso humano obtenido a partir de cultivo celular en presencia de suero humano o extracto de embrión de pollo (Isern, 2013); d) membrana de retículo endoplásmico asociada a mitocondria (MAM) de ratones con bloqueo (KO, *knockout*) de Cav-1, o tipo silvestre (*wild type*) de ratón (Wieckowski, 2009); e) lisado total de fibroblastos a partir de ratones con bloqueo de Zmpste24 o tipo silvestre (Varela, 2005); f) extracto total de corazón porcino tras una hora de reperfusion pos-infarto (Danielsen, 2007; Garcia-Prieto, 2014); g) linfocitos de ratón con bloqueo de HDAC6 y tipo silvestre de bazo de ratón (Gonzalez-Granado, 2014); y h) células de músculo liso

vascular primarias (VSMC) tratadas con angiotensina-II (AngII); en estas muestras (VSMC) se utilizaron seis preparados independientes. En el artículo se detallan las hipótesis nulas utilizadas como referencia para determinar la significatividad, las aleatorizaciones y los tests estadísticos.

El artículo que aquí se presenta ha sido preparado en estrecha colaboración con Fernando García-Marqués, que es, conjuntamente, primer autor. Mi contribución se presenta en la primera parte, y consiste en i) la modificación utilizada del algoritmo WSPP, ii) el Algoritmo de Integración Genérico (GIA), iii) el *software* que lo pone en práctica (SanXoT), iv) el Triángulo de la Biología de Sistemas (SBT) y v) el análisis bioinformático de los ocho modelos especificados en el párrafo anterior. La segunda parte del artículo trata una aplicación de gran relevancia de mi trabajo, en la que se realiza un experimento de secuencia temporal (*time-course*) con proteómica cuantitativa de alto rendimiento utilizando marcaje iTRAQ 8-plex para analizar el efecto a corto plazo de la angiotensina-II (AngII) en las células de músculo liso vascular (VSMC) mencionadas. El algoritmo detecta con éxito alteraciones funcionales en proteínas causadas por la acción coordinada del proteoma, detectando a su vez proteínas cuyo cambio es atípico comparado con el de otras clasificadas en las mismas categorías, aportando así información relevante para la interpretación biológica. La aplicación del algoritmo SBT a los datos cuantitativos sirvieron para mostrar que la AngII daba lugar a respuestas de coordinación creciente en las proteínas de las células de VSMC; se observó que dichas respuestas eran reproducibles en el tiempo. Estos resultados son de gran interés para la identificación de objetivos terapéuticos en el tratamiento de enfermedades cardiovasculares y, más específicamente, tienen implicaciones mecanicistas en la comprensión del rol de la AngII en remodelado vascular. En esta introducción no se presentan más interpretaciones biológicas de estos resultados (que están disponibles en la segunda parte del artículo) por ser parte de la tesis doctoral de Fernando García-Marqués.

En conclusión, el algoritmo presentado tiene propiedades únicas capaces de desvelar la visión de conjunto de los cambios que tienen lugar en un proteoma, especialmente al mostrar las rutas metabólicas y de señalización observables en experimentos binarios de proteómica cuantitativa, mejorando en gran medida los resultados proporcionados por otras estrategias a disposición actualmente, tales como el análisis de enriquecimientos en redes (*enrichment network analysis*) o la anotación de clases funcionales (*functional class scoring*).

A Novel Systems-Biology Algorithm for the Analysis of Coordinated Protein Responses Using Quantitative Proteomics*[S]

Fernando García-Marqués‡¶, Marco Trevisan-Herraz‡¶, Sara Martínez-Martínez‡, Emilio Camafeita‡, Inmaculada Jorge‡, Juan Antonio Lopez‡, Nerea Méndez-Barbero‡, Simón Méndez-Ferrer‡, Miguel Angel del Pozo‡, Borja Ibáñez‡, Vicente Andrés‡, Francisco Sánchez-Madrid‡, Juan Miguel Redondo‡, Elena Bonzon-Kulichenko‡§, and Jesús Vázquez‡§

The coordinated behavior of proteins is central to systems biology. However, the underlying mechanisms are poorly known and methods to analyze coordination by conventional quantitative proteomics are still lacking. We present the Systems Biology Triangle (SBT), a new algorithm that allows the study of protein coordination by pairwise quantitative proteomics. The Systems Biology Triangle detected statistically significant coordination in diverse biological models of very different nature and subjected to different kinds of perturbations. The Systems Biology Triangle also revealed with unprecedented molecular detail an array of coordinated, early protein responses in vascular smooth muscle cells treated at different times with angiotensin-II. These responses included activation of protein synthesis, folding, turnover, and muscle contraction – consistent with a differentiated phenotype—as well as the induction of migration and the repression of cell proliferation and secretion. Remarkably, the majority of the altered functional categories were protein complexes, interaction networks, or metabolic pathways. These changes could not be detected by other algorithms widely used by the proteomics community, and the vast majority of proteins involved have not been described before to be regulated by AngII. The unique capabilities of The Systems Biology Triangle to detect functional protein alterations produced by the coordinated action of proteins in pairwise quantitative proteomics experiments make this algorithm an attractive choice for the biological interpretation of results on a routine basis. ***Molecular & Cellular Proteomics 15: 10.1074/mcp.M115.055905, 1740–1760, 2016.***

From the ‡Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain

Received September 30, 2015, and in revised form, February 9, 2016

Published February 18, 2016, MCP Papers in Press, DOI 10.1074/mcp.M115.055905

Author contributions: J.R., E.B., and J.V. designed research; F.G., M.T., S. Martínez-Martínez, E.B., and J.V. performed research; F.G., M.T., E.B., and J.V. analyzed data; F.G., M.T., E.B., and J.V. wrote the paper; F.G., E.C., I.J., J.L., N.M., E.B., and J.V. performed the proteomics experiments; M.T. developed the software package; S. Méndez-Ferrer prepared biological samples; M.d., B.I., V.A., F.S., and J.R. prepared the biological samples.

Cellular processes are executed by proteins working together in complexes or functional pathways, and recent evidence indicates that proteins carry out their functions in a coordinated manner. The concept of coordinated behavior of genes and proteins has been studied from different experimental and conceptual perspectives. High throughput mRNA analysis has revealed that genes that are functionally linked or are associated with the same metabolic pathway are often co-expressed (1–3), and that multiprotein complexes within the same functional class are regulated in the same direction (4). By combining gene expression information with data about protein interactions, growth phenotype, and transcription factor binding, it was possible to depict modules or groups of genes showing correlated behavior (5). Similarly, fluorescence techniques have revealed that transcript levels of temporally induced genes are highly correlated in individual yeast cells (6). At the protein level, high-throughput single-cell flow cytometry has been used to study biological noise (7), showing that proteins that are subunits of the same complex tend to attain similar levels in the cell, that fluctuations in protein levels tend to be smaller within large complexes (8), and that cell-to-cell variability in protein expression is similar for proteins sharing a similar function (9). Using a yeast fusion library for immunodetection and measurement of absolute expression levels (10), Carmi *et al.* showed that interacting proteins are expressed in significantly more similar cellular concentrations (11).

Recent advances in mass spectrometry (MS)-based proteomics allow the identification and relative quantification of thousands of proteins in a single study, making MS the technique of choice for the study of cellular processes on a proteome-wide scale. This technology has revealed that proteins with similar functions typically have similar expression

levels (12). Similarly, a tendency of functionally related proteins to be coordinately regulated was shown by correlation analysis of protein abundance after density fractionation (13), over a time course (14), or in a large set of diverse conditions (15). Despite these efforts, the mechanisms by which cells coordinate the levels of functionally related proteins are still poorly understood. Furthermore, existing methods to analyze quantitative data obtained from conventional proteomics studies (for instance when only two different conditions are compared) are not designed to detect coordination and least of all to analyze its extent.

Here we present the Systems Biology Triangle (SBT)[1], which to our knowledge is the first algorithm able to detect coordinated protein behavior in high-throughput pairwise quantitative proteomics experiments. We show the utility of the algorithm by revealing clearly coordinated protein behavior in biological models of varied origin subjected to different kinds of perturbations. We have also applied the model to characterize the molecular alterations that take place along time when vascular smooth muscle cells (VSMC) are treated with angiotensin-II (AngII). VSMC are a highly plastic cell type with a pivotal role in vascular wall remodeling, a pathophysiological process underlying prevalent cardiovascular diseases like hypertension and aneurism (16). Although numerous lines of evidence support the implication of AngII in this process (17–20), little is known about the signaling pathways and molecular mediators involved, particularly in the early phases of the process, where the responses can be more attributable to the direct action of AngII. In this work, we have performed the first proteome-wide, time-course study of the early response induced by AngII on VSMC. Application of the SBT algorithm to the quantitative data revealed that AngII produced on VSMC an array of increasingly coordinated protein responses that were reproducibly observed along time. Unlike other approaches widely used by the proteomics community, including enrichment, network analysis, and functional class scoring methods, SBT was able to detect the coordinated activation of protein synthesis, folding, turnover, and muscle contraction machineries, consistent with a differentiated phenotype, as well as the induction of migration and the repression of cell proliferation and secretion. Our work is the first to uncover the global picture of the changes that take place in VSMC in response to AngII stimulation at early times, and reveal the regulation by AngII of a considerable number of molecular pathways that have not been previously described. Thus, our work has important mechanistic implications in understanding the role of AngII in vascular remodeling. These results show that the new algorithm has unique capabilities to detect functional protein alterations produced by the coordinated action of proteins in relevant biological systems.

[1] The abbreviations used are: SBT, Systems Biology Triangle; VSMC, Vascular Smooth Muscle Cells; AngII, Angiotensin II; GIA, Generic integration algorithm; QC, Protein to category; CA, Category to all (or grand mean); ORA, Over-representation; FCS, Functional class scoring.

EXPERIMENTAL PROCEDURES

The Generic Integration Algorithm (GIA)—

Rationale—During the analysis of quantitative proteomics data by mass spectrometry, the quantitative information is obtained from the spectra and this information is used to quantify the peptides from which the spectra are produced and the proteins that generate these peptides. In other words, the quantitative information is integrated from the spectrum level to the peptide level and then from the peptide level to the protein level. Here we first describe a generic integration algorithm (GIA) (Fig. 1*A*) that can be used to construct compact workflows containing sequential integration steps for the analysis of quantitative data (Fig. 1*B*) and the detection of significant protein abundance changes. Afterward, we show that GIA can be applied to construct workflows containing the Systems Biology Triangle (SBT) (Fig. 1*C*), a kind of integration that allows the detection of changes in functional categories produced by the coordinated behavior of their proteins.

Formulation of the GIA Model—Let us assume that the abundance of two samples is being compared and that during the process of quantification we are rolling up the quantitative information from a lower level (L) (*e.g.* peptide level) to calculate the quantitative values at a higher level (H) (*e.g.* protein level). We will use the scheme "*L* to *H*" to refer to this integration step (*e.g. peptide to protein* integration). We define the set of values x_l as the relative abundances of the species in the lower level expressed as the 2-base logarithm of the ratio of abundance of the species present in each sample, and the set of values x_h as the relative abundances of the species in the higher level (Fig. 1*A*). We express the relationships between the elements of the lower and higher levels by a relations table that indicates the correspondence between the indexes *l* and *h* of the elements of the lower and higher levels (Fig. 1*A*).

The model assumes that in the null hypothesis the quantification of any element *l* at the lower level may be expressed in the form of a mixed-effects model in relation to the quantification of the element *h* at the higher level it belongs to:

$$x_l = x_h + \rho_{hl} + \beta_l,\ l \in h \qquad \text{(Eq. 1)}$$

where ρ_{hl} is the random deviation between the element of the lower level and that from the higher level (*e.g.* the error introduced when peptides are prepared from their corresponding proteins), β_l contains the random error carried by the element *l* of the lower level because of previous integrations from lower levels (*e.g.* the error carried by the peptide values as they are previously determined from the average of several spectra), and $l \in h$ indicates that *h* is the element from which *l* is derived (*e.g.* that the peptide comes from a given protein).

The model assumes that the random errors are normally distributed:

$$\rho_{hl} \sim N(0, \sigma^2_{LH})$$

$$\beta_l \sim N(0, \sigma^2_l)$$

The parameter σ^2_{LH} *is the general variance of the L to H* integration, which is constant for all species upon which the integration is performed. Note that this assumption implies that all elements at the lower level are affected by the same variance in relation to the higher level. For instance, in the *peptide to protein* integration, all the peptides are assumed to be affected by the same error source when they are generated from the corresponding proteins.

The parameter σ^2_l *is the prior local variance, and corresponds to the variance that affects to element l* of the lower level because of previ-

ous integrations (*e.g.* the variance that affects to any peptide because of the quality of quantifications at the spectrum level). We define the prior weight, v_l, as the inverse of σ_l^2.

If we define the statistical weight w_{lh} as the inverse of the total variance of x_l, we will have

$$w_{lh} = \frac{1}{\sigma_l^2 + \sigma_{LH}^2} = \frac{1}{\frac{1}{v_l} + \sigma_{LH}^2} \quad \text{(Eq. 2)}$$

When the *L* to *H* integration is performed, the quantitative values at the higher level are calculated as weighted averages of the values at the lower level:

$$x_h = \frac{\sum w_{lh} x_l}{\sum w_{lh}},\ l \in h \quad \text{(Eq. 3)}$$

According to error propagation theory, the inverse of the variance of the average is the sum of the inverses of the variances of the averaged elements. Therefore, the prior weights of the elements at the higher level, v_h are the sum of statistical weights of the corresponding elements at the lower level:

$$v_h = \sum w_{lh},\ l \in h \quad \text{(Eq. 4)}$$

Finally, the standardized variables of the elements of the lower level corresponding to the *L* to *H* integration are calculated as

$$z_{lh} = (x_l - x_h)\sqrt{w_{lh}}\sqrt{\frac{n_h}{n_h - 1}},\ n_h > 1 \quad \text{(Eq. 5)}$$

where n_h is the number of elements from the lower level that are integrated to calculate x_h. The values of z_{lh} are expected to distribute according to the standard normal distribution N (0, 1). Note that although z_{lh} is not defined when $n_h = 1$, this does not preclude calculation of x_h and v_h.

Integrative Workflow and Estimation of the General Variance of the Integration—In the GIA integration the prior weights of the elements of the higher level are calculated from those in the lower level. However, the prior weights of the lower-level elements in the first integration step (*e.g.* the elements at the spectrum level) are not defined; these weights must be calculated in advance using other algorithms. In this work, we estimated the variances of the elements at the spectrum level using the method described in a previous work (21).

The input of the GIA algorithm is the table (id_l, x_l, v_l) at the lower level, in plain tab-delimited text format, where id_l is the identifier of the element *l*, and the relations table (id_h, id_l) that establishes the correspondence of the identifiers in the lower level with those in the upper level (Fig. 1*A*). The output is the table (id_h, x_h, v_h) at the higher level and the general variance of the *L* to *H* integration, σ_{LH}^2. The algorithm also computes the statistical weights at the lower level, w_{lh}, and the standardized variables of the lower level corresponding to the *L* to *H* integration, z_{lh}.

It is important to note here that the variables w_{lh} and z_{lh}, which fully describe the distribution of errors at the lower level, can only be calculated once the *L* to *H* integration is performed. In consequence, the output does not contain the complete information about the variance of the higher level. For instance, in the case of the *peptide to protein* integration, the total variance of each peptide, and therefore the standardized variable at the peptide level, cannot be estimated until peptides are integrated into proteins. In addition, this integration does not give information about total protein variance, which remains unknown until proteins are integrated at higher levels. Therefore, in this computational scheme the variances are indissolubly associated with the integrations.

In each integration, σ_{LH}^2 is estimated using a robust, iterative method. An initial seed value of $\sigma_{LH}^2 = 0.001$ is used by default, and the statistical weights at the lower level w_{lh} are calculated using eq. 2. The integration is then performed to obtain the quantifications x_h at the upper level using eq. 3, and the standardized variables of the integration z_{lh} calculated using eq. 5. Once these parameters are calculated, the difference

$$\Delta = abs\left|\frac{median(z_{lh}^2)}{\{\Phi^{-1}(3/4)\}^2} - 1\right| \quad \text{(Eq. 6)}$$

where $\Phi^{-1}(x)$ is the inverse error function, is used as a robust parameter to determine the deviation of the z_{lh} distribution from the standard normal distribution $N(0,1)$. The process is iterated to minimize Δ using the Levenberg-Marquardt algorithm (22).

Determination of Confidence Intervals of the Variance—Confidence intervals of the variance were calculated using an algorithm that creates simulated experiments. As input, it used the original set of (x_l, v_l) values, and as output it produced an altered ($x_{l,random}$, v_l) data set, where x_l is substituted by a random error generated according to the variance of x_l:

$$x_{l,random} = x_l + \frac{\Phi^{-1}(rand(0,1))}{\sqrt{w_{lh}}} \quad \text{(Eq. 7)}$$

where *rand*(0, 1) is a random decimal number between 0 and 1. For each experiment, 21 simulated experiments were generated, and the variance σ_{LH}^2 was calculated for each one by applying the GIA. After sorting the resulting list of variances, the median (11th element), along with the 3rd highest and 3rd lowest variances were taken to get the closest values to a one-sigma confidence for each side, in accordance to

$$\|21 \cdot (1 - \Phi(1))\| = 3$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

Estimation of FDR and Detection and Elimination of Outliers—The false discovery rate (FDR_{lh}) is calculated by ranking the data according to the absolute values of z_{lh} (from high to low) and estimating the expected number of false deviations from the null hypothesis:

$$FDR_{lh} = \frac{2 \cdot (1 - \Phi(|z_{lh}|))}{rank(|z_{lh}|)/N} \quad \text{(Eq. 8)}$$

where N is the total number of z_{lh} values. Note that in the plots presenting the distributions of z_{lh} the values are ranked from low to high.

Outliers in the *L* to *H* integration are defined as the elements at the lower level whose FDR_{lh} surpasses a given threshold (typically 1% or 5%). Outliers can be automatically eliminated by an iterative method. An initial GIA is executed to calculate FDR_{lh} and σ_{LH}^2. Then, for each element at the higher level, the most extreme lower level outlier is removed, and the GIA is repeated maintaining the original σ_{LH}^2 variance. The procedure is repeated until no further outliers are detected.

The Systems Biology Triangle—In biological systems proteins can be classified as belonging to a functional class or category, meaning by that any general form of classification, including molecular function, cellular component, pathway, or interaction network. We hypothesized that if during the perturbation of a biological system all the proteins belonging to a category undergo the same abundance change, then the quantitative behavior of the set of proteins can be described by that of their category. We also hypothesized that the relative abundance of the category can be estimated by integrating the quantitative values of its protein components, using the GIA algorithm.

To develop an algorithm to detect significant changes in abundance of functional categories produced by the coordinated behavior of their protein components, we designed the Systems Biology Triangle (SBT). The SBT is constructed by applying the GIA algorithm to integrate from *protein to category* and then from *category to all*, as schematized in Fig. 1*C* (blue triangle). The variances calculated in the SBT analysis provide different levels of information. In the ideal case of complete coordination all the proteins distribute normally within each category, so that there are no protein outliers in the categories and therefore z_{qc} follows a perfect N (0, 1) distribution (Fig.1*D*, left). In this case, the variance σ^2_{QC} of the *protein to category* integration only reflects the experimental error by which the proteins are quantified around the corresponding category average, *i.e.* the experimental protein variance. In such a situation, the biological perturbation should only affect the abundance of the categories, and this effect would become evident in the *category to all* integration when the distribution of z_{ca} is confronted against the null hypothesis that the categories are not affected by the perturbation, *i.e.* $\sigma^2_{CA} = 0$ (Fig. 1*D*, left). In the extreme case of a noncoordinated behavior the perturbation affects proteins irrespective of their category, and therefore σ^2_{QC} would have a higher value than in the case of full coordination (Fig. 1*D*, right). Similarly, the perturbation should have no effect on the distribution of categories (Fig. 1*D*, right), which would distribute according to $\sigma^2_{CA} = 0$. In this situation we would not detect significantly changing categories. We speculated that in a real case the situation would be intermediate between these two extremes.

Therefore, we designed the SBT algorithm as follows: firstly, the *protein to category* integration is performed and the GIA-built in algorithm is applied to remove protein oultiers in each category. By eliminating the protein-category outliers we assured that the abundance change of the proteins remaining in the category did not significantly deviate from the category average; we defined such a situation by stating that these proteins follow a *coordinated behavior*. The categories that are significantly altered by the perturbation as a consequence of the coordinated behavior of their proteins are then detected as outliers in the *category to all* integration, by analyzing the distribution of z_{ca} values under the null hypothesis that $\sigma^2_{CA} = 0$. We also defined the *degree of coordination* of the biological response as the fraction of nonoutlier (*i.e.* coordinated) proteins belonging to the changing categories in the population of proteins that change individually or form part of changing categories (Fig.1*E*).

SBT analysis can also improve the detection of individual protein abundance changes. These changes are detected through the analysis of the *protein to all* integration (Fig.1*B*). However, the variance σ^2_{QA} obtained at this stage reflects the total deviations of protein quantifications around the grand mean, which includes not only the experimental uncertainties by which proteins are quantified, but also, to some extent, the effect of the perturbation on protein abundance. In proteomics publications it is generally assumed that the majority of the proteins do not change at the time of estimating protein variance; however, in the absence of a null hypothesis, it is impossible to determine whether this assumption is able to capture the experimental variance in a distribution influenced by a biological perturbation. We speculated that the variance σ^2_{QC}, estimated through the SBT analysis (Fig. 1*C*) would be less influenced by the perturbation than σ^2_{QA}, and therefore would constitute a more precise, conservative estimate of the experimental protein variance.

Clustering Algorithm—To facilitate the detection of similar categories (categories sharing many proteins), a clustering algorithm was applied. The set of changing categories was represented as a weighted directed graph, called *similarity graph*, where each vertex was associated to a category. Edges between two categories were weighted according to the number of proteins shared. A minimal threshold for the weights was calculated by a Durfee square-based optimization algorithm (23), and edges having a weight below the threshold were removed. This procedure left a number of isolated subgraphs containing categories that share similar sets of proteins, allowing the detection of the main category describing the behavior of the group.

Experimental Design and Statistical Rationale—The SBT model was tested by analyzing a total of eight different biological models subjected to quantitative proteomics by several stable isotope labeling methods. These included: 1) yeast cells treated with H_2O_2 or with vehicle (Yeast/H_2O_2); 2) untreated yeast cells (Yeast/-) (21); 3) human bone marrow mesensphere secretome obtained from cells cultured in the presence of human serum or chicken embryo extract (Human Secretome/HS *versus* CE) (24); 4) liver mitochondrial-associated membranes from Cav-1 knockout or wild type mice (Mouse MAM/Caveolin-1 KO) (25); 5) total fibroblast lysates from Zmpste24$^{-/-}$ or wild-type mice (Mouse Fibroblasts/Zmpste24 KO) (26); 6) total heart extracts from pigs after 1 h post-infarction reperfusion (27, 28), in these samples the infarcted and remote areas of the same animal were compared (Pig Heart/Infarct); 7) mouse lymphocytes obtained from HDAC6 KO and wild-type mouse spleen (Mouse Lymphocytes/HDAC6 KO); in these samples activated total spleen cells were cultured in 10% FBS-RPMI with IL-2 stimulation (29); and 8) primary mouse vascular smooth muscle cells (VSMC) treated with angiotensin-II (AngII) (see below); in these samples six independent biological preparations were used. Details about the null hypothesis used as a reference to determine significance, randomization and statistical tests used are described above. Stable isotope labeling and peptide identification and quantification methods are described below.

Vascular Smooth Muscle Cell Culture—Primary vascular smooth muscle cells were isolated from C57BL/6 mouse abdominal and thoracic aortas and cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 20% fetal bovine serum (FBS) (30). VSMC were rinsed thrice with PBS and starved for 48 h in serum-free DMEM. Cells were than treated with 1 μM of AngII (Sigma-Aldrich) or vehicle in serum-free culture medium for 2, 4, 6, 8, and 10 h. Cells were lysed in RIPA buffer (NaCl 150 mM, 50 mM Tris pH 8.0, 1% Nonidet P-40, 0.5% sodium desoxycholate, 0.1% SDS) supplemented with 1 mM DTT, 3 mM EGTA, 1 mM PMSF, and protease inhibitor mixture (Sigma-Aldrich). Protein extracts were stored at −80 °C until use.

Sample Preparation, Peptide Labeling and Fractionation, and Mass Spectrometry Analysis—Yeast cells treated with H_2O_2 or with vehicle (Yeast/H_2O_2) and different technical replicates of untreated yeast cells (Yeast/-) were treated and labeled with SILAC, ^{18}O or iTRAQ and the resulting data integrated as described (21). Trypsin digestion of all protein extracts was performed using a robust protocol described previously (31). VSMCs and total lymphocyte mouse lysates were labeled with iTRAQ8-plex and 4-plex, respectively, according to manufacturer's instructions. The secretome of human mesenspheres, murine liver MAMs, pig heart extracts, and mice total fibroblast lysates were subjected to ^{18}O labeling (31).

Labeled peptides were subjected to peptide separation into six fractions using a Waters Oasis MCX cartridge (Waters Corp, Milford, MA). Briefly, desalted and dried peptides were taken up in 1 ml of 5 mM ammonium formate pH 3 (AF3). Each MCX cartridge was equilibrated by slowly passing 1 ml of 1:1 methanol/water across cartridge, followed by 3 ml of AF3 containing 25% acetonitrile (ACN) (v:v). Sample was applied at flow rate of 1 drop per second and cartridge was washed with 1 ml of AF3–25% ACN. The bound peptide was eluted into six fractions with 1 ml of freshly prepared buffers: (1) 500 mM AF3, 25% ACN, (2) 1 M AF3, 25% ACN, (3) 1.5 M AF3, 25% ACN, (4) 500 mM AF3, 1.5 M potassium chloride (KCl), 25% de ACN, (5) 1.25 M AF3, 37.5% ACN, and (6) 1 M AF3, 50% de ACN. Eluted peptides were dried in speed vac, desalted on HLB OASIS cartridges (32), and stored at −20 °C until MS analysis.

The samples were analyzed using LC-MS instrumentation consisting of an Easy nano-flow HPLC system (Thermo Fisher Scientific) coupled via a nanoelectrospray ion source (Thermo Fisher Scientific) to either an LTQ-Orbitrap Elite for ^{18}O-labeled samples or to a Q Exactive mass spectrometer for iTRAQ-labeled samples (both Thermo Fisher Scientific). For LC, C18-based reverse phase separation was used with a 2-cm trap column and a 50-cm analytical column (EASY column, Thermo). Peptides were loaded in buffer A (0.1% formic acid (v/v)) and eluted with a 360 min linear gradient of buffer B (90% acetonitrile, 0.1% formic acid (v/v)) at 200 nL/min. Mass spectra were acquired in a data-dependent manner, with an automatic switch between MS and MS/MS using a top 20 method. MS spectra were acquired in the Orbitrap analyzer with a mass range of 390–1500 *m/z* and 120,000 resolution at *m/z* 400 (Orbitrap Elite) or 390–1500 *m/z* and 35,000 resolution at *m/z* 200 (Q Exactive). CID peptide fragments, acquired at 30 of normalized collision energy (Orbitrap Elite) or HCD peptide fragments obtained at 25 of normalized collision energy (Q Exactive), were analyzed at high resolution in the Orbitrap.

Peptide Identification and Quantification—The raw files were analyzed with Proteome Discoverer (version 1.4, Thermo Fisher Scientific), using a Uniprot database containing all human and chicken protein sequences (November 23th, 2011; 47,609 entries) for the human mesenspheres samples; all mouse protein sequences (April 28th, 2012; 122,974 entries) for the mouse samples; a joint Human+Yeast Swissprot database (Uniprot release 57.3 May 2009; 26,885 entries) for the yeast samples, and a joint Human+Pig Swissprot database (May 30th, 2012; 153,506 entries) for the pig samples. For database searching, parameters were selected as follows: trypsin digestion with two maximum missed cleavage sites, precursor mass tolerance of 800 ppm, fragment mass tolerance of 50 mmu. For ^{18}O-labeled samples variable methionine oxidation, lysine and arginine modification of +4 Da and fixed cysteine carbamidomethylation were used. For iTRAQ, we allowed variable methionine oxidation and fixed cysteine carbamidomethylation, lysine and N-terminal modification of +144.1020 Da for iTRAQ 4-plex or + 304.2054 for iTRAQ 8-plex. The same collections of MS/MS spectra were also searched against inverted databases constructed from the same target databases. Peptide identification from MS/MS data was performed using the probability ratio method (33). False discovery rates (FDR) of peptide identifications were calculated using the refined method (34, 35); 1% FDR was used as criterion for peptide identification. Each peptide was assigned only to the best protein proposed by the Proteome Discoverer algorithm. Quantitative information was extracted from MS spectra, for ^{18}O samples, or MS/MS spectra, for iTRAQ samples, using an in-house developed program (QuiXoT), as described (21), and protein abundance changes were analyzed using the Generic Integration Algorithm, as described above. Calculation of statistical weights of each quantitation at the spectrum level was performed according to the WSPP model (21). The validity of the null hypothesis at each one of the levels (spectrum, peptide, protein, and functional category) was carefully checked by plotting the cumulative distributions, as described (21). iBAQ parameter for all quantified proteins was calculated as in (36).

Protein Functional Annotation—Quantified proteins were functionally annotated using Ingenuity Knowledge Database (IPA) (37, 38), CORUM (39), and DAVID (40). The latter repository included 15 functional databases, such as KEGG, REACTOME, Gene Ontology, and Panther, among others. The Qiagen Transcription Factors Database was taken from http://www.sabiosciences.com/chipqpcrsearch.

Western Blotting and Immunofluorescence—Whole-cell lysates were prepared by lysing cells with RIPA buffer (50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 3 mM EGTA and 1% Nonidet P-40, 0.5% sodium desoxycholate, 0.1%SDS) supplemented with 1 mM DTT, 1 mM PMSF, and a mixture of protease inhibitors (Sigma-Aldrich) on ice for 30 min and then boiling in 2× Laemmli buffer. Lysates were subjected to SDS-PAGE followed by immunoblotting with antibodies against various proteins, including calponin (Santa Cruz Biotechnology), tenascin (Millipore, AB19011), thrombospondin-1 (Thbs1) (Thermo Scientific, MS-421-B0), Prostaglandin G/H synthase 2 (PTGS-2) (Cayman, 160126) and methionine adenosyltransferase IIβ (Mat2β) (Santa Cruz Biotechnology, sc-390586). Tubulin-α (TUBA) (Sigma-Aldrich, T 6074) was used as protein loading control.

For immunofluorescence of cultured VSMCs, cells seeded onto coverslips were subjected to AngII treatment (10^{-6} M) for different times, fixed with 3% paraformaldehyde, permeabilized with 0.1% Triton-X-100, and then stained with anti-Calponin (1/100; Abcam 46794), or antitype III collagen (1/100, Abnova MAB1514), followed by Alexa Fluor 488-labeled secondary antibody. Nuclei were stained with Hoechst. Slides were mounted and visualized using an inverted confocal microscope (LSM700; Carl Zeiss) with 25× oil objective. Images were processed for presentation with Zen 2012 software (Carl Zeiss).

Data Access—The software needed to execute GIA workflows can be downloaded from ftp://ftp.cnic.es/ftpsvc/pub/SanXoT_package_Source_Code_Example.zip.

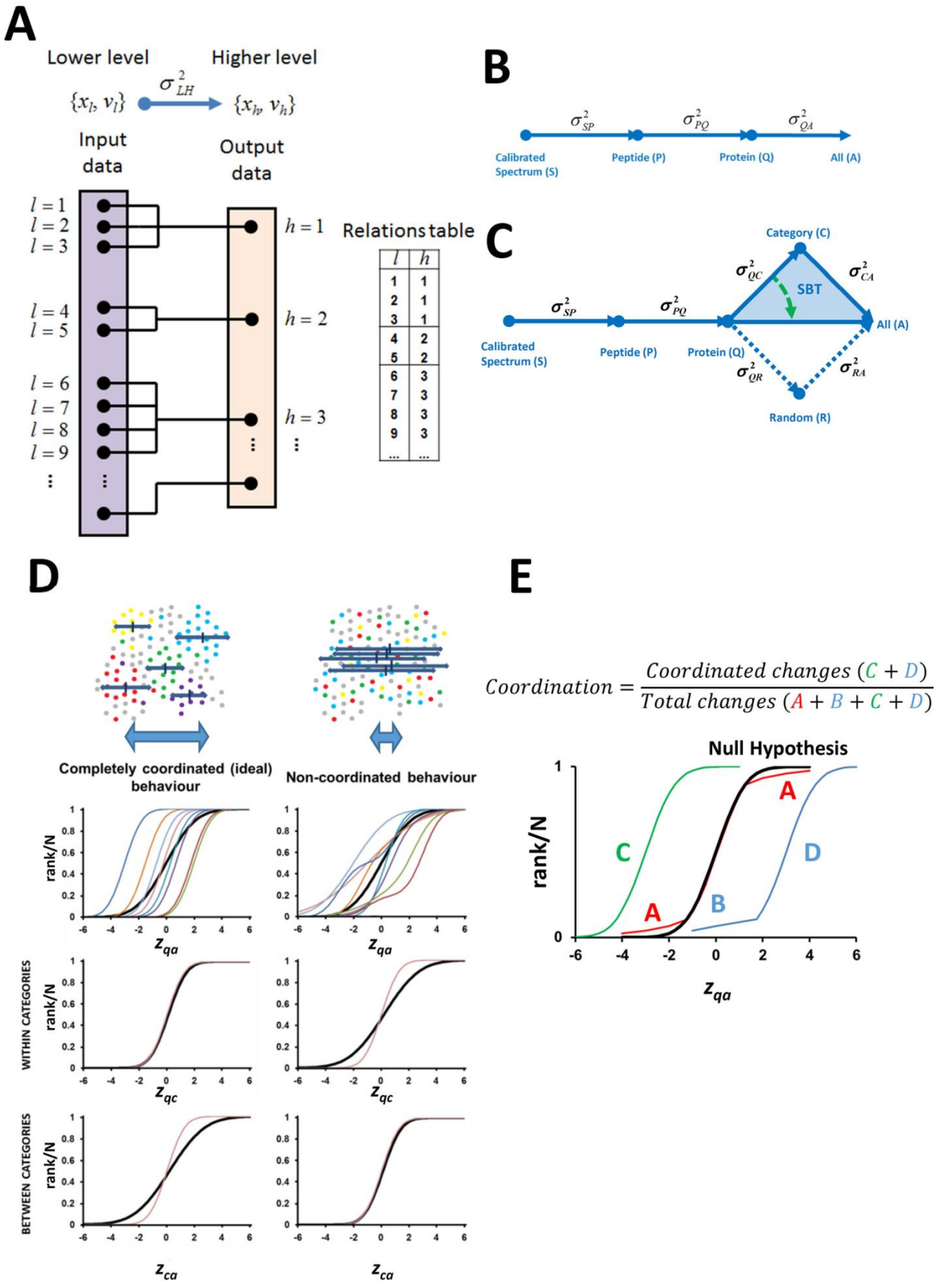
A readme.txt file is provided with basic instructions to install and execute the package. Help for each program is displayed by using the *-h* parameter.

The data set from the yeast experiment (21) is available in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (41) under data set identifier PXD000325. The data set from the analysis of VSMC proteome (raw and msf files and excel table with identification and quantification data) is available in the PeptideAtlas repository (http://www.peptideatlas.org/PASS/PASS00690) that can be downloaded via ftp.peptideatlas.org, username: PASS00690, password: PU2454dpa.

RESULTS

A Generic Integration Algorithm to Analyze Quantitative Proteomics Data Allows the Development of the Systems Biology Triangle, a Functional Class Scoring Algorithm That Captures Protein Coordination—We firstly developed a generic integration algorithm (GIA) to integrate quantitative data from a lower to a higher level using error propagation theory (Fig. 1*A*) (see Experimental Procedures). By applying GIA sequentially, we developed automated workflows for conventional analysis of pairwise quantitative proteomics experiments (Fig. 1*B*), in which significant protein abundance changes were detected as outliers in the *protein to all* integration step.

We then applied GIA to construct the Systems Biology Triangle (SBT), a workflow that performs the *protein to category* integration, eliminates protein outliers within each category and detects significant category abundance changes produced by the coordinated behavior of their proteins (Fig. 1*C*). In parallel SBT detects significant protein abundance changes. The rationale of SBT design is explained under "Experimental Procedures." In the most general case, biological systems will be somewhere between two extreme situations (Fig. 1*D*): a complete coordination (left), where the effect of the perturbation is noticed in the *category to all* integration (CA), and a completely noncoordinated behavior (right), where both the experimental error and the effect of the perturbation are reflected at the protein (QC) level. We defined the *degree*


A
Lower level
Higher level
σ^2_{LH}
$\{x_l, v_l\}$
$\{x_h, v_h\}$
Input data
Output data
$l=1$
$l=2$
$l=3$
$l=4$
$l=5$
$l=6$
$l=7$
$l=8$
$l=9$
$h=1$
$h=2$
$h=3$
Relations table
l h
1 1
2 1
3 1
4 2
5 2
6 3
7 3
8 3
9 3
B
σ^2_{SP}
σ^2_{PQ}
σ^2_{QA}
Calibrated Spectrum (S)
Peptide (P)
Protein (Q)
All (A)
C
Category (C)
σ^2_{QC}
σ^2_{CA}
SBT
σ^2_{SP}
σ^2_{PQ}
All (A)
Calibrated Spectrum (S)
Peptide (P)
Protein (Q)
σ^2_{QR}
σ^2_{RA}
Random (R)
D
Completely coordinated (ideal) behaviour
Non-coordinated behaviour
rank/N
z_{qa}
WITHIN CATEGORIES
z_{qc}
BETWEEN CATEGORIES
z_{ca}
E
$Coordination = \frac{Coordinated\ changes\ (C + D)}{Total\ changes\ (A + B + C + D)}$
Null Hypothesis
rank/N
C
A
A
B
D
z_{qa}

of coordination of the biological response as a measure of the fraction of proteins that are coordinated (Fig. 1*E*).

SBT Reveals Coordinated Protein Behavior in Real Biological Systems—To test the SBT model in practice, we first reanalyzed a published data set describing a quantitative comparison of untreated (A) and 0.5 mM H_2O_2-treated (B) *Saccharomyces cerevisae* cells (Yeast/H_2O_2), obtained by integrating the information from eight different replicate experiments using a variety of stable isotope labeling approaches (21). From that study, we also used a parallel integration of eight additional experiments where replicates from nontreated samples were compared as the null hypothesis (Yeast/-). Application of the SBT to the Yeast/- experiment produced *z*-variables that followed the *N*(0, 1) distribution in both the QC and CA integrations (Fig. 2*A*-1), with no significant category changes, indicating excellent agreement between experiment and theory. In contrast, clear evidence of coordination was found in the Yeast/H_2O_2 experiment, showing deviations from the null hypothesis at the category level (CA integration) with almost negligible deviations in proteins within categories (QC integration) (Fig. 2*A*-2). In addition, the proportion of protein-category outliers was remarkably low (1.6%, supplemental Table S1). To compare these results with those that would be obtained in a completely noncoordinated behavior, we randomly shuffled the quantitative data assigned to each protein without altering the *protein to category* relations table, so that the original relational structure and the protein redundancy in the categories were maintained; the process was repeated several times to determine average values and confidence intervals. Randomization increased the variance of proteins within their categories (compare σ^2_{QR} with σ^2_{QC}) and caused category variance to disappear (compare σ^2_{RA} with σ^2_{CA}) (Fig. 2*B*), showing that the coordination captured by the SBT was statistically significant. The pattern of coordinated protein alterations was clearly evident from analysis of protein changes within each category (Fig. 2*C*-1 and 2).

SBT analysis of the category alterations in yeast revealed clear internal coherence in the coordinated protein responses induced by H_2O_2 (supplemental Fig. S1 and supplemental Table S2). Thus, there was a general repression of biosynthetic metabolic pathways, showing that oxidative stress did not selectively alter only a minority of metabolic enzymes (21), but rather produced a profound metabolic reconfiguration. Consistently, other processes central to cellular homeostasis were also decreased, such as sulfur compound biosynthesis, cell growth and protein trafficking, and proteasome activation was consistent with the induction of a cellular response for rapid elimination of misfolded oxidized proteins (42). Finally, the rest of activated categories where consistent with active mitochondrial adaptation to adverse conditions (43), cellular restoration of metal homeostasis (44, 45), and replacement of ribosomal proteins and rRNA damaged by oxidative stress (46).

SBT was then used to analyze a collection of five high-throughput quantitative experiments performed in several biological models in response to various stimuli and using different stable isotope labeling methods. As shown in Fig. 2*A* 3–7, in all cases the distributions of z_{qc} and z_{ca} indicated the presence of a coordinated behavior, with clear deviations at the category level and almost negligible deviations of proteins within categories. Moreover, the fraction of proteins that were detected as outliers in their categories was remarkably low in all situations (supplemental Table S1). Similarly, randomization of protein values increased protein variance and almost eliminated category variance (Fig. 2*B*). Finally, the distributions of abundance changes (z_{qa} values) of proteins within the most representative changing categories (Fig. 2*C*) revealed that in all the biological models most categories were displaced along the *x* axis while maintaining the same sigmoidal shape, as expected for a coordinated behavior.

In a further study, we analyzed whether the database used for ontological protein classification influenced the degree of coordination detected by our model. We selected three mouse experiments and repeated the SBT analysis using several functional classification databases. The degree of coordination and the fraction of coordinated categories tended to be higher when using curated databases like PANTHER and KEGG and had a tendency to diminish when using databases containing more general terms, such as GO and GO Slim (supplemental Fig. S2). However, the absolute number of coordinated categories tended to increase, because of the higher information content of the latter ones, suggesting that the optimum choice is a compromise between sensitivity and specificity.

Protein Coordination During VSMC Activation with AngII Builds Up Dynamically Over Time—We next analyzed whether the SBT model was able to detect coordinated protein abun-

FIG. 1. **Generic integration algorithm (GIA) for quantitative proteomics and extension to the Systems Biology Triangle (SBT).** *A*, Scheme with the elements of the GIA, where $\log_2$-ratios of species in the lower level (x_l) are rolled up taking into account their quantification weights (v_l) to obtain the relative abundances of the species in the higher level (x_h), (21). The relations table between the elements of the lower (*l*) and higher levels (*h*) is also schematized. *B*, Workflow for the analysis of a quantitative proteomics experiment using GIA. *C*, Workflow for the systems biology analysis of a quantitative proteomics experiment including the Systems Biology Triangle (SBT) (shaded in blue). Note that the variance obtained in the *protein to category* integration can be used as improved estimation of the experimental protein variance (green arrow). *D*, Scheme of the concept of proteome coordination according to the SBT, where the behavior of a hypothetical, completely coordinated proteome response (left panel) is compared with that of a completely noncoordinated response (right panel). The black curves represent the null hypothesis. *E*, Definition of the degree of coordination used in this work. The letters A and B represent the proteins outliers in the red and blue categories, whereas C and D represent the coordinated proteins in the green and blue categories.

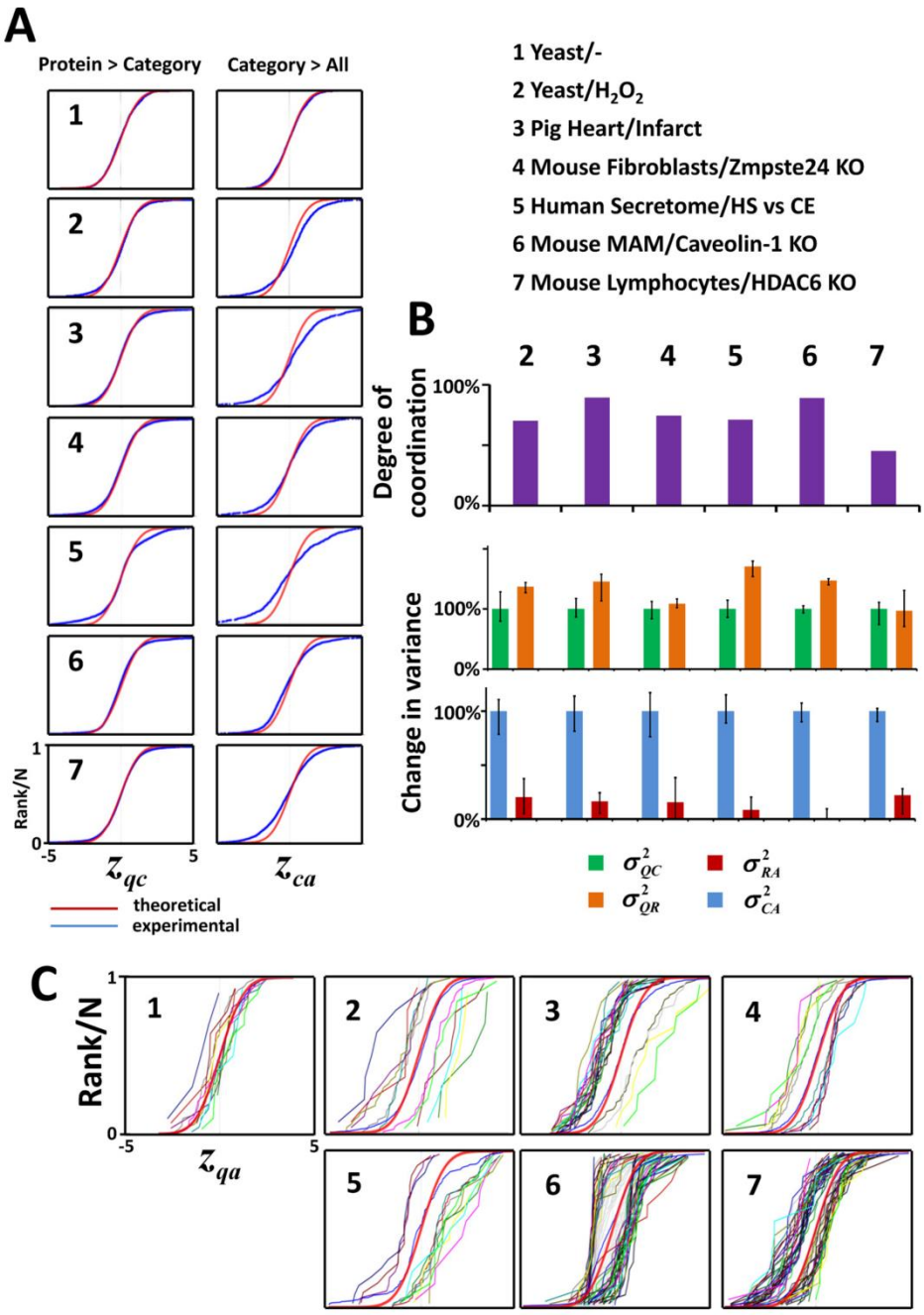


FIG. 2. **Evidence of coordinated protein behavior in real biological systems.** *A*, The distribution of the standardized variables describing the quantification variability among proteins within their category (z_{qc}) and among different categories around the grand mean (z_{ca}), resemble more a coordinated than a noncoordinated behavior in all proteomes analyzed. The proteomes analyzed are described in Methods. *B*, Degree of coordination and effect of randomization on the variances. Randomization increases the protein-category variance and diminishes the category-all variance to values that are not significantly different from zero. Data are expressed as median ± standard deviation. *C*, Distribution curves of the standardized variable of coordinated proteins (z_{qa}) plotted separately in the different changing categories.

dance changes over time when primary VSMC were stimulated with AngII, a factor involved in vascular wall remodeling (47, 48). Cells were incubated with AngII for 0, 2, 4, 6, 8, and 10h and the quantitative changes were analyzed using multiplexed isobaric labeling followed by peptide fractionation to increase proteome coverage. This experiment also served to test if coordination of the same proteins could be reproducibly observed in different biological preparations from the same cell type.

Analysis of protein abundance changes was performed by applying the GIA-derived workflow (Fig. 1*B*), obtaining standardized values that in all the cases and integration levels followed normal distributions (supplemental Fig. S3 and supplemental Table S3*A*). The results revealed a reproducible pattern with a slight tendency to increase over time (Fig. 3*A*). We then analyzed coordinated protein changes using SBT (Fig. 1*C*). The quantitative protein information was integrated into functional categories using a database of more than 4500 categories constructed from IPA and DAVID repositories and also protein complexes from CORUM. 95% of the proteins could be annotated using these databases. Remarkably, the proportion of *protein to category* outliers was very low, and <3% in all cases (supplemental Table S3*B*), and no evidence of coordination was detected when the data were subjected

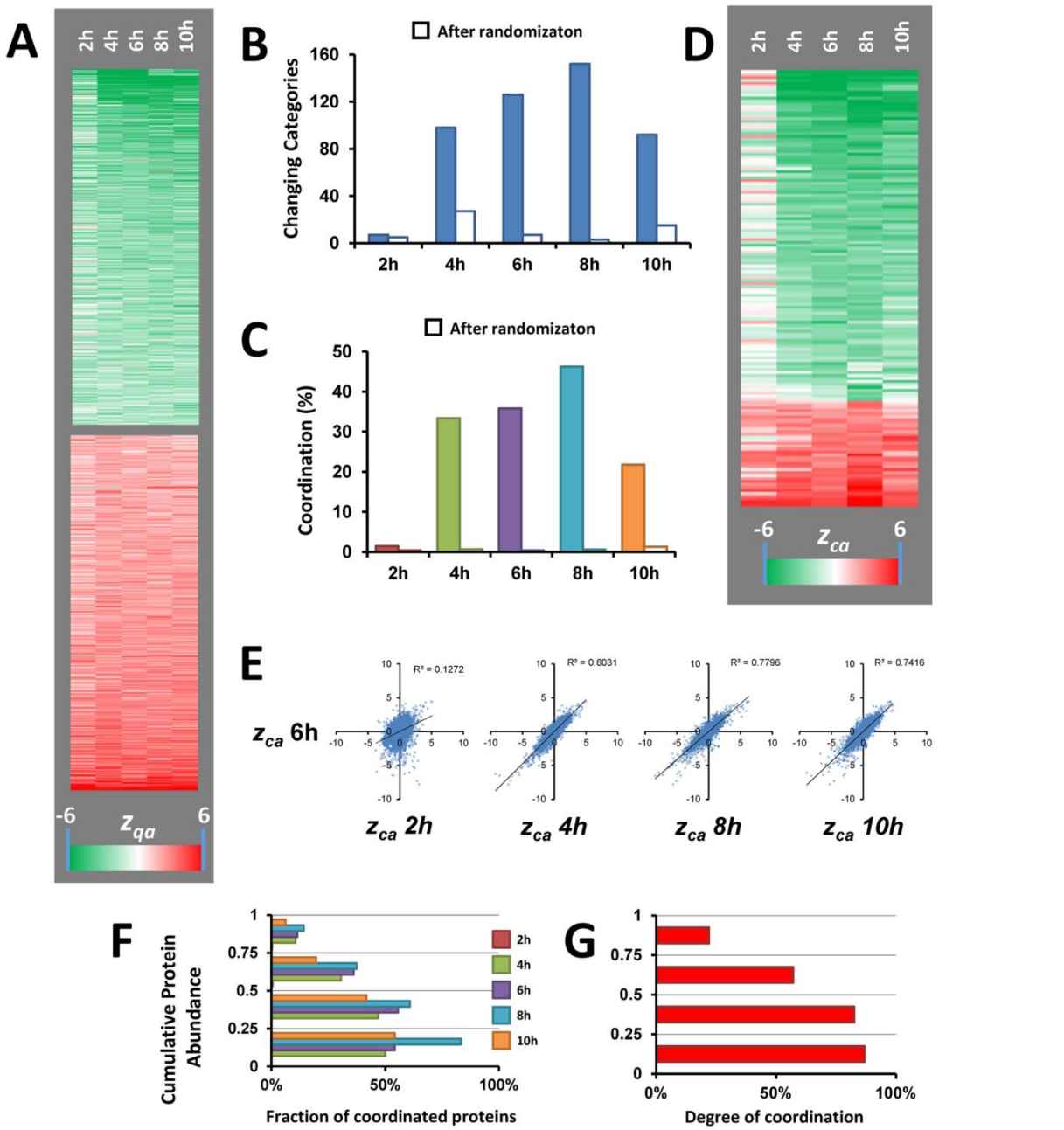


FIG. 3. **Analysis of protein coordination in the VSMC proteome along the time course of AngII-treatment.** *A*, Heat-map showing the time-course of the standardized protein quantifications (z_{qa}) of the 1000 proteins having the most significant abundance change. The magnitude of the standardized variable is shaded according to the color scale at the bottom. *B*, Time-course of the number of categories changing in a coordinated manner. *C*, Time-course of the degree of coordination. *D*, Heat-map showing the time-course of the standardized category quantifications (z_{ca}) of the categories changing in at least one time-point. The magnitude of the standardized variable is shaded according to the color scale at the bottom. *E*, Correlation analysis between the standardized category quantifications at the different time-points in relation to those at 6h of AngII treatment. *F*, Distribution of coordinated proteins in the four iBAQ abundance quartiles at the different time points, in relation to the total number of proteins in each quartile. *G*, Distribution of protein coordination within the four iBAQ abundance quartiles. Coordination was calculated considering together all the proteins in the five time points.

to randomization (Fig. 3*B* and *C*). Interestingly, after removal of outliers, a clear pattern of coordinated category changes emerged, but only after 4 h of AngII treatment (Fig. 3*B–E*). Correlation analysis revealed that the categories altered at 4, 8, and 10 h, but not those at 2 h, were essentially the same that were altered at 6 h (Fig. 3*E*). Consistently, at 2 h practically all proteins changed in a noncoordinated way and hence were detected as category outliers, and these independent changes were similar along time (supplemental Fig. S4). These results indicate that proteins begin to change as early

as 2h, being this change predominantly uncoordinated, and that coordinated behavior is gradually built up at later times. We also observed that the proportion of proteins changing in a coordinated manner tended to increase with abundance (Fig. 3*F*), and the majority of proteins in the quartiles of higher abundance were coordinated (Fig. 3*G*). These findings may suggest that a tight regulation of the most abundant proteins may be required upon stimulation by AngII to maintain cell homeostasis; however, this interpretation should be taken with caution, because the lower abundance proteins have a higher protein variance, making it more difficult to detect coordinated protein behavior.

AngII Promotes Coordinated Protein Changes Consistent with a Hypertrophic and Contractile VSMC Phenotype—Because the analysis used several functional category databases, and categories were often redundant in their protein content, to simplify interpretation we developed a clustering algorithm that assembled changing categories containing similar proteins into functional clusters, so that a representative category could be defined for each cluster (supplemental Fig. S5). This step simplified the outcome, reducing the total number of AngII-sensitive protein categories that changed in a coordinated manner to 25 (supplemental Table S4); after this process only the collagen and muscle protein categories needed a further classification (supplemental Fig. S6 and supplemental Table S5; see also below). Interestingly, more than one third of the categories corresponded to known protein complexes described in CORUM, in good agreement with the expected coordinated behavior of complex-forming proteins (11, 49, 50).

A subset of these complexes were involved in DNA packing and mRNA splicing, and included the nucleosome core, composed by histones, which constitute the basic scaffold for DNA packing (51), and small nuclear ribonucleoproteins, which form part of the spliceosome (Fig. 4). Other complexes were involved in ribosomal protein synthesis, including the 40S and 60S subunits of ribosome, and in protein folding and degradation, including the prefoldin complex, the 20S core particle of proteasome and cyclophilins, which have been reported to form part of spliceosomes (52). Together these categories depict a clear picture of coordinated activation of protein synthesis, folding and turnover machineries (Fig. 4).

We also detected a coordinated increase of a group of categories that were all associated with a contractile and differentiated phenotype (Fig. 5), including muscle proteins, tropomyosins and other well-known smooth muscle components (SM22/calponin-related proteins) (53). Among these, we observed a coordinated increase in proteins composing the B-Ksr1-MEK-MAPK-14-3-3 complex, which acts as a scaffold integrating the actions of diverse proteins including kinases, transcription factors, and apoptotic molecules (54). This complex translocates to the plasma membrane in response to growth factors, regulating Ras signaling, and producing several outcomes in VSMC, ranging from contraction to proliferation (55). We also detected an increase in abundance of the 5HT3 type receptor mediated signaling pathway, which promotes calcium release from intracellular stores, causing smooth muscle contraction (56). An increase in abundance of calcium-activated proteins involved in the regulation of PGC-1α was also in clear agreement with the augmented contractile capacity of VSMC, because PGC-1α promotes biogenesis of mitochondria (57), an important intracellular calcium reservoir and ATP source for muscle contraction. Accordingly, we observed an increase in the protein abundance of most of the components of mitochondrial F1F0-ATP synthase. Interestingly, a STRING-based interaction analysis revealed that all these groups of proteins formed highly interconnected networks, which sometimes joined together several of these categories (Fig. 5). This array of physical and functional protein interactions might be the cause of the high degree of coordination between the proteins contained in these categories, and explains why these categories were captured by the SBT model.

AngII Induces Coordinated Protein Alterations Reflecting VSMC Migration and Repression of Cell Proliferation and Secretion—An additional group of coordinated categories was related to the induction of VSMC migration (Fig. 6). Among these, the cytoskeletal regulation by RhoGTPase and ARF/SAR proteins were both upregulated by AngII. RhoGTPases act as the primary Ca^{2+} sensitizers in smooth muscle, increasing the contractile output of these cells (58), and participate in the formation of focal adhesions, which are important for cell migration (59, 60). Proteins belonging to the ARF/SAR category have important roles actin remodeling (61, 62) and some of its members act in coordination with Rac1 and RhoA (63), which are implicated in cell migration. AngII-activated small GTPases have also been implicated in promoting migration in neuronal and epithelial cells (64, 65). Further evidence toward a migratory phenotype was provided by the coordinated upregulation of proteins from the Arp2/3 complex, a classical podosome component of smooth muscle cells (66), and the Upstream Regulator_MAP3K1 category. The components of this category include tenascin, a glycoprotein providing de-adhesive properties to the cells and enabling cell movement (67), and thrombospondin-1, a potent stimulus for VSMC migration (68). These changes were paralleled by downregulation of fibrillar collagen, which is known to suppress smooth muscle cell migration (69). Finally, the upregulated VSMC migration response is consistent with a progressive increase upon AngII treatment of the levels of PTGS-2 (supplemental Table S6), a well-known VSMC migration marker (70). Interestingly, the six most AngII-responsive proteins from the ARF/SAR category are regulated by the transcription factor PAX-4a (according to the Qiagen Transcription Factors Database), whereas the proteins of the rest of categories form complexes or densely interconnected networks (Fig. 6), providing a molecular basis for the detection of coordinated behavior by the SBT model.

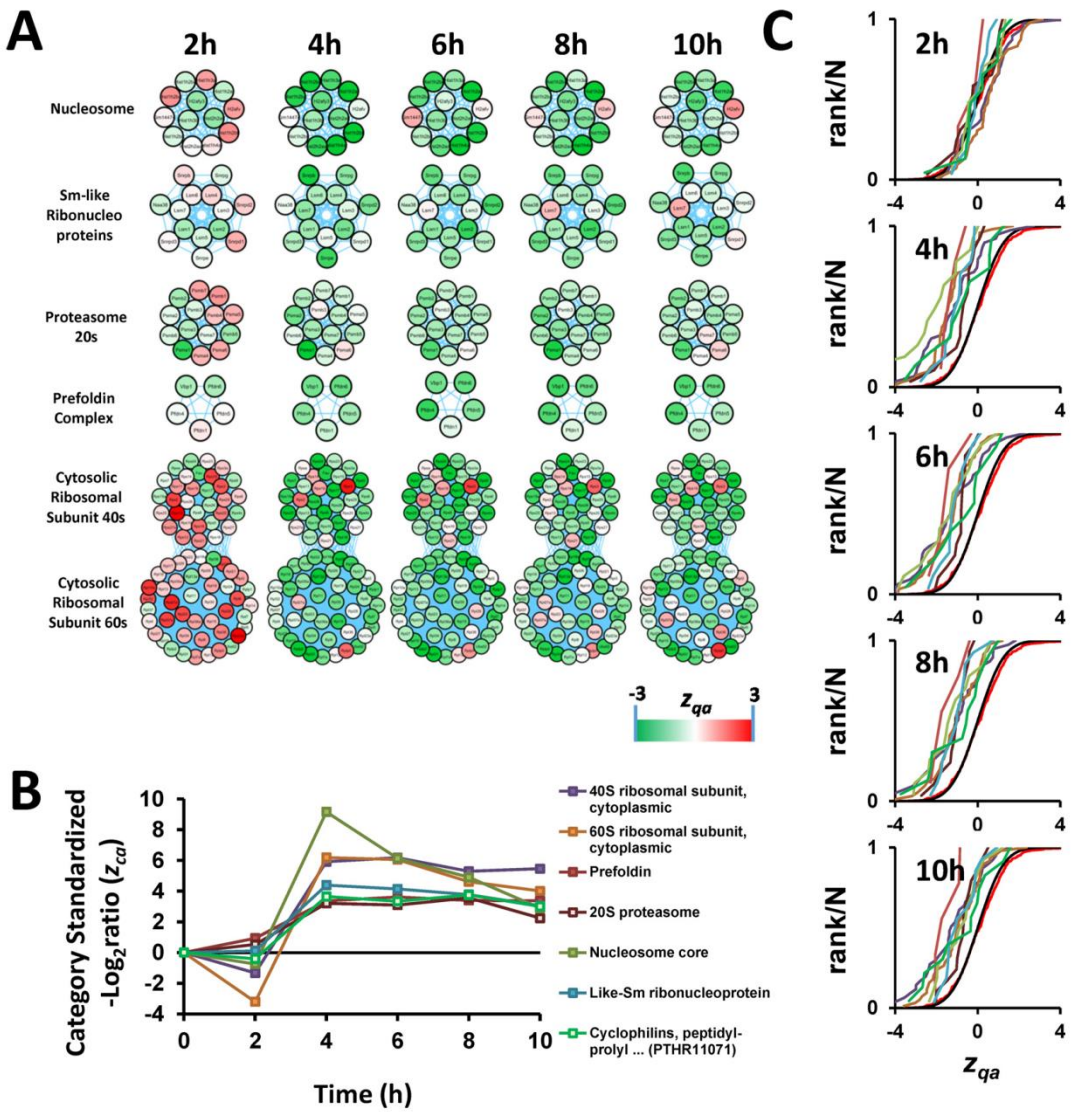


FIG. 4. **AngII promotes a coordinated activation of protein synthesis, folding, and turnover in VSMC.** *A*, Time course of protein changes belonging to categories implicated in protein synthesis, folding, and turnover machineries. All these proteins are known to form protein complexes. Although it is not clear which specific histone isoforms interact to each other to form the nucleosome, we represented them as a complex because it is well known that eight of these proteins (two copies of each histone protein, H2A, H2B, H3, and H4) bind together to form the core of the nucleosome (124). In addition, the proteins annotated as "Sm-like ribonucleoproteins" are actually from three different complexes LSm1–7, LSm2–8, and SMN (125, 126), and all of them are part of the spliceosome. *B*, Time-course of changes in the categories shown in panel *A*. For the sake of graph clarity, the z_{ca} of categories that increase in response to AngII are shown as positive. *C*, Distribution of the standardized protein quantifications (z_{qa}) belonging to the categories indicated in the legend of panel *B*, showing coordinated protein changes.

Finally, we detected a coordinated alteration of several categories indicating a repression in cell proliferation and secretion (Fig. 7). Thus, we detected repression of the DNA-dependent ATPase MCM complex, which in eukaryotes is composed by six proteins (MCM2 to 7) and is required for initiation of chromosome replication (71, 72), and the glutamine amidotransferase category, containing mostly key enzymes involved in glutamine use for RNA and DNA synthesis (73), were both repressed (Fig. 7*A*). Homocysteine biosynthesis was also decreased (Fig. 7*B*). This pathway promotes smooth muscle cell proliferation (74) and is an important methylation route targeting DNA (75, 76), mRNA (75), and histones (77, 78). This methylation reaction is catalyzed in part by members of the protein arginine methyltransferase (Prmt) family, four of which were downregulated (Fig. 7). Histone methylation generally promotes tighter DNA packing around nucleosomes, resulting in heterochromatin formation and repression of gene expression (79). Repression of this route would thus promote more relaxed chromatin packing and consequent histone release from nucleosomes, which agrees well with the observed increased in histone levels (Fig. 4). The complement pathway, known for triggering inflammation (80)

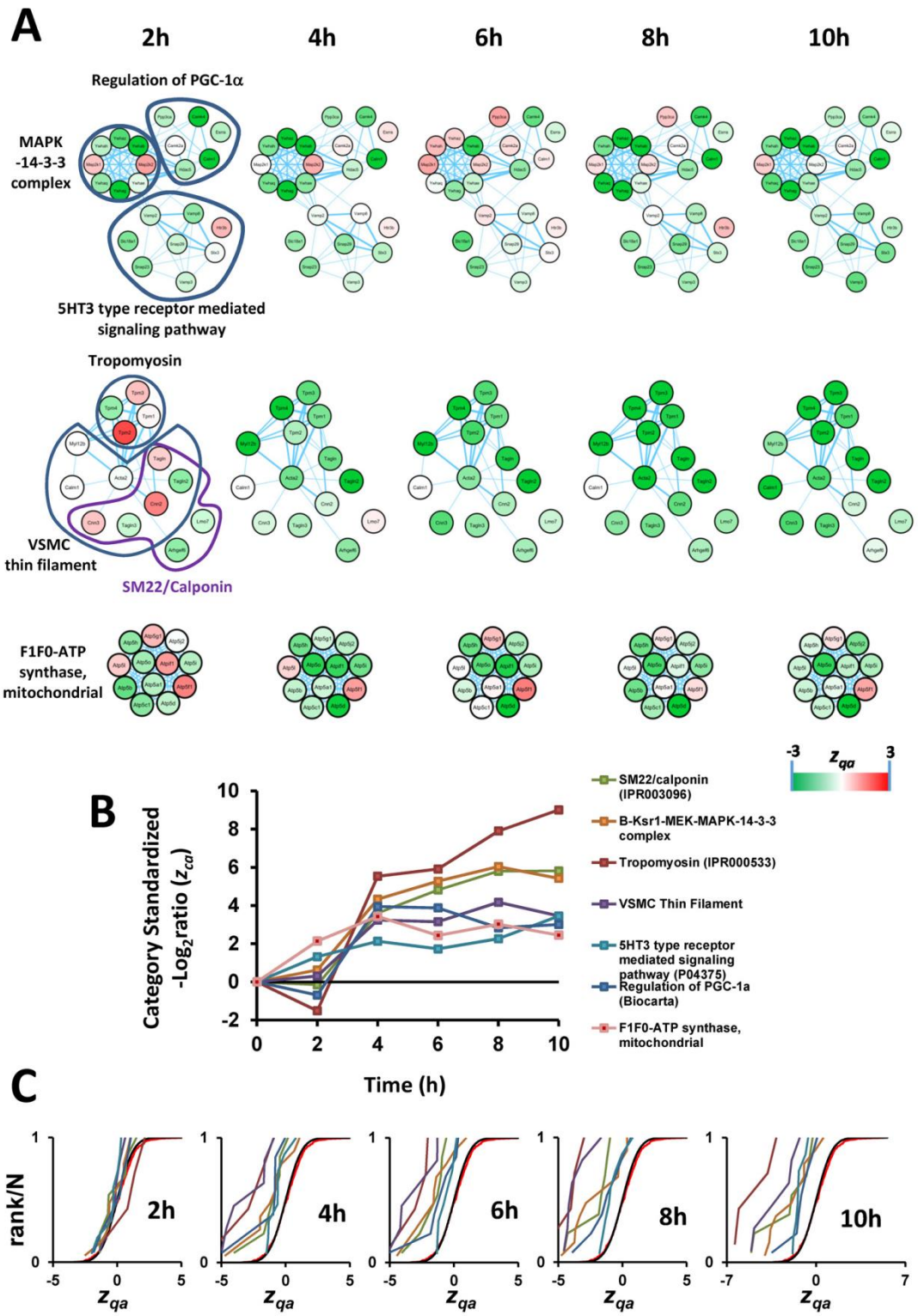


FIG. 5. **Evidences of coordinated protein changes consistent with a contractile VSMC phenotype upon AngII stimulation.** *A*, *B*, and *C*, have the same meaning as in Fig. 4. Known protein complexes are represented as in Fig. 4; the rest of proteins are represented as an interaction network according to String.

and VSMC switch to a proliferative phenotype (81), was also downregulated (Fig. 7*D*). In addition, antioxidant enzymes and gluconeogenesis were coordinately upregulated (Fig. 7*A*). The coordinated upregulation of gluconeogenesis, which was detected together with a concomitant repression of glycolytic enzymes (Fig. 7*C*), is consistent with nonproliferative pheno-

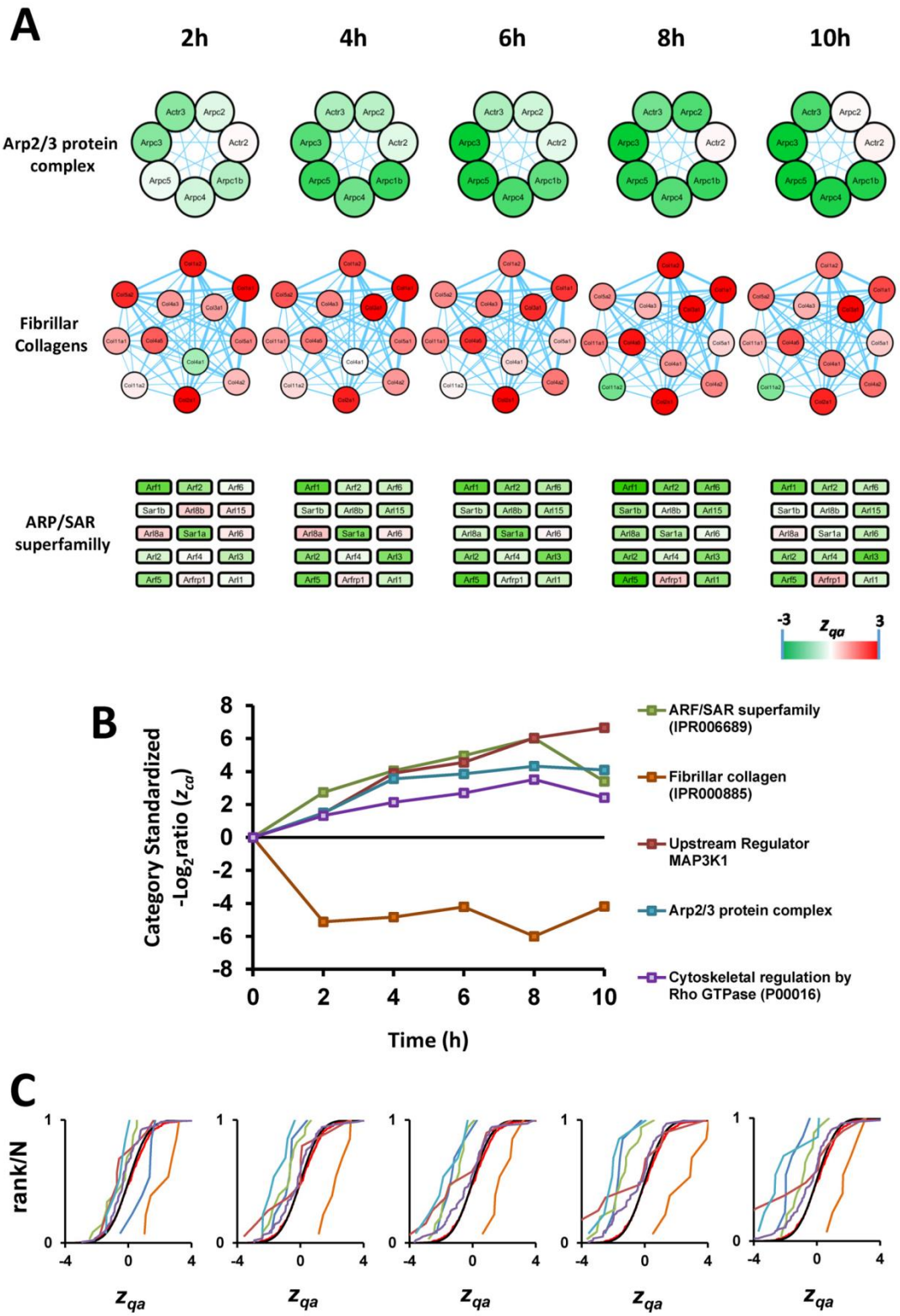


FIG. 6. **AngII induces coordinated protein alterations reflecting VSMC migration.** *A*, *B*, and *C*, have the same meaning as in Fig. 5. Proteins not known to form complexes or not forming a String interaction network are presented as squares.

type of VSMC because proliferative VSMC display a highly glycolytic phenotype (82). Similarly, the increase in antioxidant enzymes would reflect the activation of defense mechanisms. This activation would counteract the described action of AngII as a promoter of oxidative stress in VSMC (83), which has been described as a mediator for VSMC proliferation (84).

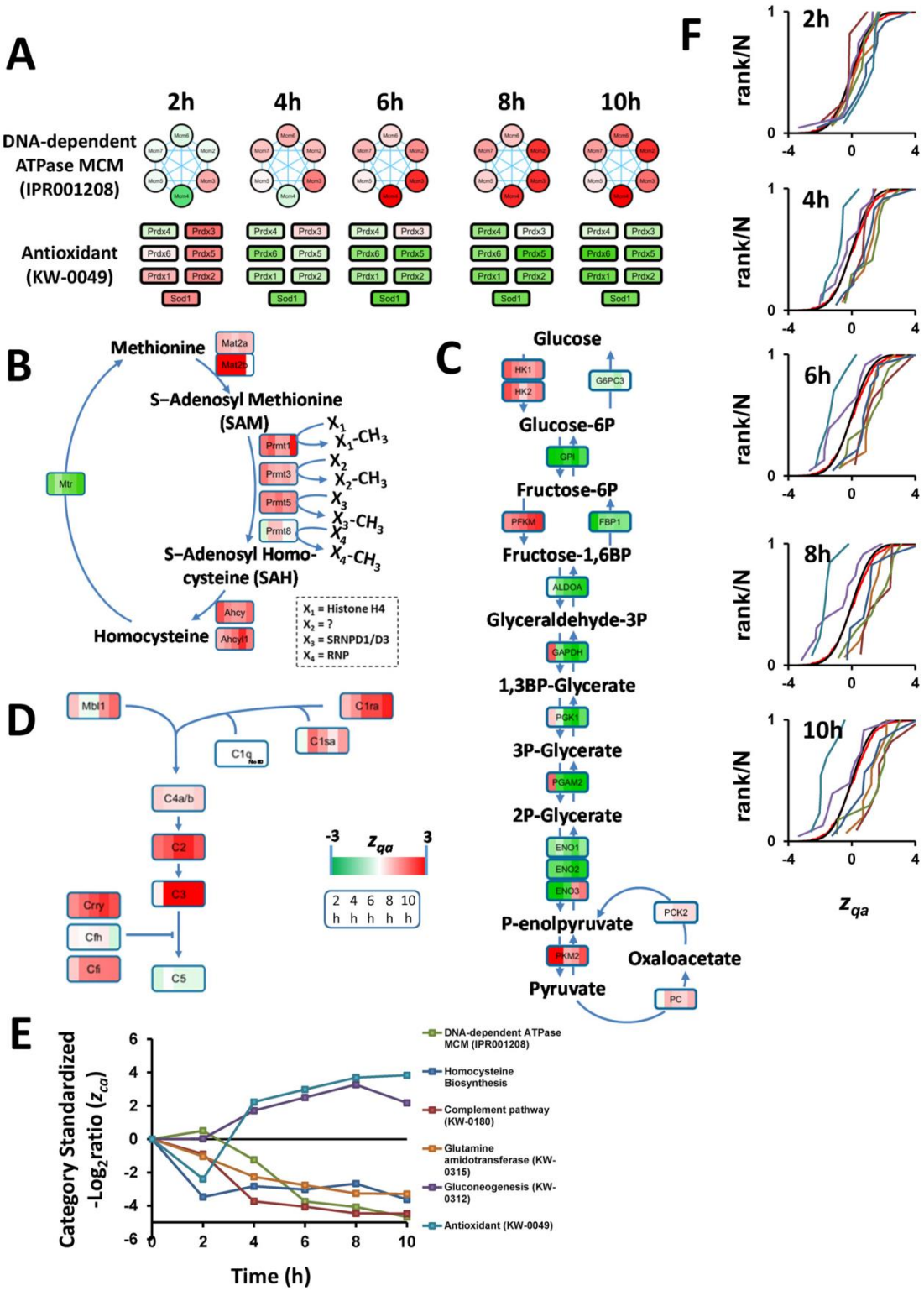


FIG. 7. **AngII induces coordinated protein alterations reflecting VSMC repression of cell proliferation and secretion.** *A*, *E*, and *F*, have the same meaning as in Fig. 5. *B*, *C*, and *D*, Schematic representation of protein changes in the homocysteine biosynthesis, gluconeogenesis, and complement pathways, respectively. Each protein is shown as a colored square, where each stripe represents the standardized protein quantification at each time point.

Remarkably, with the exception of MCM proteins, which were detected as a CORUM complex, within each category most proteins related to repressed cell proliferation and secretion were regulated by the same transcription factors. For example, the most responsive enzymes from the homocysteine pathway were regulated by NFkappa-B (85), a key mediator of AngII-driven proliferation and migration (86). Moreover, antioxidant enzymes are known to be regulated through antioxidant-response elements (87, 88) recognized by the nuclear factor (erythroid-derived 2)-related factor 2 (Nrf2) (89). Similarly, the complement system genes share common effectors, such as IL-1, IL-6, and interferon gamma (90). Gluconeogenesis enzymes also share common activating transcription factors, including the CREB-regulated transcription coactivator 2 and the glucocorticoid receptor (91). Together these results reveal that most coordinated categories that are discovered by the SBT model include proteins that share either physical or functional interconnections or common mechanisms of regulation.

Additional technical approaches were used to confirm the progressive acquisition of the contractile, nonproliferative, and migratory VSMC phenotype in response to AngII. Immunofluorescence staining detected increased levels of the contractile protein calponin (CNN) and decreased levels of type-III collagen (Fig. 8*A*), whereas Western blot revealed reduced levels of methionine adenosyltransferase 2β (Mat2β), the key regulatory enzyme of the homocysteine biosynthesis pathway (Fig. 8*B*). Western analysis also showed increased abundance several migration-associated factors, including tenascin (TnC), thrombospondin-1 (Thbs1), and prostaglandin-endoperoxide synthase 2 (PTGS-2) (70, 92–94) (Fig. 8*C*).

Protein Outliers Maintain Their Noncoordinated Behavior Over Time and Tend to Play Differential Functional Roles—The functional category alterations detected by our model are not affected by the presence of isolated proteins (outliers) showing high changes in abundance, because these outliers are not considered to analyze the behavior of categories. This raises the question whether these outliers have specific biological roles that explain their abnormal behavior. To answer this question we first analyzed individual protein changes and separated those proteins belonging to a changing category (as described above). Remarkably, independent proteins (that change but do not belong to a changing category) were detected as early as 2h, when coordination was not yet detectable (compare Fig. 3*C* with supplemental Fig. S4). Moreover, these independent proteins consistently maintained their behavior over time, displaying in all cases a close resemblance to the behavior at 6h (supplemental Fig. S4*B* and *C*). Outlier proteins thus seem to respond to the stimulus differently to proteins following a coordinated pattern.

In a second analysis, we inspected the nature of the clearest protein outliers that belonged to the 25 categories altered by AngII (supplemental Table S7). Some of them belonged to a subset of proteins with a clearly separated functionality,

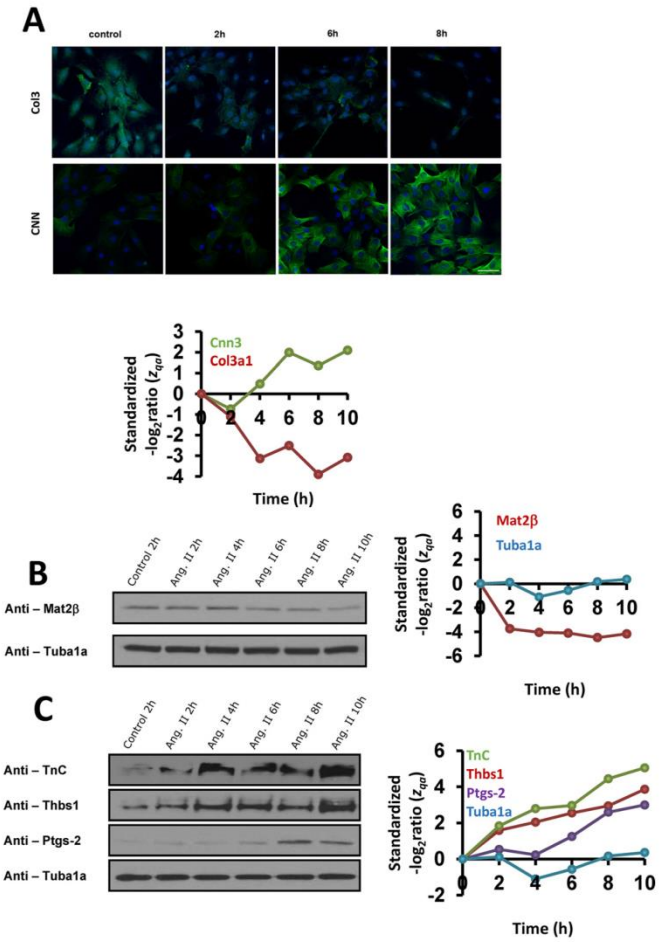


FIG. 8. **Immunological validation of the results obtained by the SBT.** *A*, Representative calponin (CNN) and type-III collagen (col3) immunostaining performed with VSMC treated with AngII for 2, 6, and 8 h. Bar, 50 μm. *B*, and *C*, Western blots showing methionine adenosyltransferase IIβ (Mat2β) downregulation and tenascin (TnC), thrombospondin-1 (Thbs1) and prostaglandin-endoperoxide synthase 2 (PTGS-2) upregulation by AngII along time. Tubulin-α (TUBA) was used as protein loading control. The time-course of the standardized protein quantification (z_{qa}) for these proteins are also indicated; for the sake of graph clarity, the z_{qa} of categories that increase in response to AngII are shown as positive values.

suggesting inappropriate classification. This was the case of some collagens and muscle proteins (supplemental Table S5 and supplemental Fig. S6). Among the remaining proteins, PROF2 and PROF1, which were opposite outliers in the same category, were found to be differentially regulated; thus, PROF1, has been described to promote cell motility, whereas PROF-2 acts as a suppressor (95). MYH10, an outlier also assigned to this category, is a nonmuscle heavy chain myosin and is known to play specialized functions in cytokinesis, concentrating in microspikes at the tips of filopodia and retraction fibers (96–98). Similarly, the ribosomal phosphoprotein P0 (Rplp0), the only outlier detected in the category of 60S ribosomal proteins, participates in protein complexes with other cytosolic proteins (Grap2 and cyclin D interacting

protein) (99), and its free form has been implicated in other extra-ribosomal functions (100). Finally, PAI-1, the clearest example of an outlier changing in the same direction as its category, is a major mediator of the effect of AngII on VSMC (101). Together these results indicate that the outliers detected by the model correspond to proteins that are either incorrectly classified or play specific roles that are different from those of the other proteins in the category.

Finally, in a third analysis, we inspected the independent proteins. From these, we selected a subset of 21 high-response (HR) proteins (supplemental Table S8). Interestingly, most of these proteins have been reported to play a relevant role in cell signaling, migration, and proliferation (supplemental Table S8), but none of them has been previously shown to be regulated by AngII in VSMC. Taken together, our results provide compelling evidence that the protein outliers detected by the SBT model are proteins with an independent behavior, probably because of their involvement in specific functions different from those of the proteins detected as having a coordinated behavior.

Results Obtained Using the SBT Model could not be Reproduced Using Other Representative Algorithms Widely Used by the Proteomics Community—To compare the performance of SBT with existing methods, we tried to interpret the quantitative proteomics results using several approaches widely used by the proteomics community. The proteins showing significant abundance changes were subjected to functional enrichment and network analysis using GOrilla (102) and STRING (103). Among the proteins that showed significant abundance changes ($FDR_{qa} < 5\%$) and were upregulated, we were only able to detected a significant enrichment in two functional categories: structural constituent of ribosome and calcium-dependent protein binding; similarly, only a significant network of functional interactions was evident among ribosomal and proteins related to calcium-binding. No significant enrichment or interaction network was detected among the proteins that had significant abundance changes and were downregulated. Similar results were obtained using GSEA (104), a functional class scoring algorithm that was not influenced by the threshold used to determine significant protein abundance changes. For GSEA analysis we used the same quantitative values at the protein level and the same functional category database used by the SBT model. Although SBT only detected coordinated categories, the sensitivity of GSEA was clearly inferior, being only able to detect 14% of the category changes detected by SBT at 5% FDR_{ca} (supplemental Fig. S7). The differences were even more remarkable for categories containing less than 100 proteins, where GSEA only detected 7% of the changes detected by SBT, whereas SBT practically detected all the category changes detected by GSEA. These results suggest that the SBT model is unique not only for its ability to capture protein coordination, but also for its sensitivity to detect functional category alterations.

DISCUSSION

In this study we present SBT, which to the best of our knowledge is the first algorithm that analyzes coordinated protein behavior in high-throughput quantitative pairwise proteomics experiments and is able to detect functional categories affected as a consequence of coordinated protein behavior. In contrast with other approaches used to study protein coordination, which are commented in the introduction, our approach is not based on correlation analysis, but takes information from ontological databases to classify proteins into functional categories. SBT directly addresses the question whether a protein has a coordinated behavior within a category by determining if it has the same relative abundance change as the other proteins in the category. Significant changes at the category level are then detected using only the population of proteins that have a coordinated behavior in each category. Consistently, we define the degree of coordination as the fraction of changing proteins that are coordinated, including in the calculation the proteins that belong to changing categories. This approach allows the analysis of protein coordination when only two conditions are compared, making it possible to measure the evolution and consistency of coordination over time. Here we show that this approach can capture coordinated protein behavior in seven different biological models subjected to various kinds of perturbations. However, in our extensive experience with the model we have successfully detected functional changes because of coordinated action of proteins in more than 40 comparative pairwise experiments. For this reason, we believe that the SBT model provides the proteomics community with a very useful tool for biological interpretation of pairwise quantitative proteomics experiments, obviating the need to calculate correlations in large sample data sets.

Because our model needs previous ontological knowledge about the quantified proteins, it is expected to be very sensitive to the quality and exhaustiveness of the classification. Indeed, we found a higher degree of coordination in databases using a more exhaustive classification method, such as PANTHER and KEGG, whereas larger, less curated databases, tended to detect a higher number of coordinated proteins and categories, but with a lower degree of coordination (supplemental Fig. S2). We also found that inclusion of information about protein complexes, such as the CORUM database, was particularly useful for detecting coordination, in agreement with the evidence that subunits of complexes have a remarkable tendency to be coregulated (8, 11, 14). Interestingly, we found that most of the outliers analyzed here seem to play a differential role, like regulatory or signaling functions, suggesting that our algorithm can be used to detect protein abundance changes of particular biological relevance. Our finding resembles results obtained in a previous report, showing that deviation from the correlation pattern by one or more

proteins indicated a specific function during the biological process (14).

Quantitative transcriptomics data is usually interpreted in relation to biological knowledge stored in databases, a procedure known as gene set analysis or pathway analysis (105). These knowledge databases can contain ontological information about genes, like GO or Kyoto Encyclopedia of Genes and Genomes (KEGG), or provide information about gene/protein interactions and how and where these occur, like Reactome or STRING. Pathway analysis algorithms that use ontological information are classified into two major subtypes: over-representation (ORA) and functional class scoring (FCS) (105). ORA, also known as enrichment analysis, statistically evaluates whether the subset of genes showing significant expression changes relative to a given threshold is enriched in a given category (ontology) (106, 107). This approach is widely used to analyze quantitative proteomics data, but algorithms of this kind only consider significantly changing proteins and therefore ignore most of the acquired quantitative information; moreover, they do not consider protein abundance or fold-change information. In FCS methods, the quantitative values of all genes from a category are integrated to produce a category-level value (105), which is statistically analyzed to determine significant category changes (104, 108–110). Although these threshold-free methods have also been used in proteomics and take account of all the information obtained, they were originally designed to treat transcriptomics data and therefore do not take optimal account of specific characteristics of protein quantification by mass spectrometry. Particularly, they do not consider the large dynamic range of protein concentrations typical of biological systems, which makes MS-based quantification of proteins present in low amounts very challenging and, in general, less reliable. This problem is aggravated by undersampling, whereby the number of peptides used to quantify a protein is variable and cannot be controlled (21). Furthermore, current FCS methods are not designed to analyze the presence and extent of coordination in protein behavior.

The algorithm presented in this work is conceptually inspired on the WSPP model, an algorithm that we developed to treat quantitative proteomics data produced by mass spectrometry using stable isotope labeling techniques (21). The SBT model introduces a generic integration algorithm that rigorously integrates errors made at the inferior level with the error made at the superior level, allowing full control of the error associated with the integrated elements (*i.e.* proteins or categories) so that quantitative information may be analyzed using the standardized *z* variable (*i.e.* abundance changes expressed in units of standard deviations). The SBT model takes into account the dynamic range of protein abundances by introducing statistical weights when performing data integration, and also addresses the problem of undersampling. The model allows the detection of *protein to category* outliers, so that analysis of categories after outlier removal allows the detection of significant category changes that are produced by proteins acting in a coordinated manner, and hence can also be used as a robust functional class scoring algorithm. These properties of SBT make this model unique to treat quantitative proteomics data and to detect functional alterations that take place in a coordinated manner. These properties are not shared by other algorithms like GOrilla, STRING, or GSEA, and this fact explains why in this work they were not as sensitive as the SBT model to detect the functional categories affected by AngII in VSMC.

We should also note here that the SBT model is based in a null hypothesis that has been experimentally shown. In previous studies we showed that the *z* variable at different levels consistently follows the expected *N* (0, 1) distribution (21, 31, 111). Here, using two lines of evidence, we further show that, in the null hypothesis, the standardized variable at the category level (z_{ca}) also follows the standard normal distribution. First, we show the absence of category changes in a null hypothesis experiment comparing two preparations of untreated yeast cells. Second, we show the absence of category changes when *protein to category* assignations were randomized, a process that disrupts any underlying coordination at the category levels. Therefore, the category changes may be detected by robust estimates of statistical significance. Notably, our integration model assigns more weight to quantifications of higher quality, making it unnecessary to eliminate poor quantifications (21, 111).

In this report, we present the first quantitative, time-course proteomics study on the effects of a vascular remodeling factor, AngII, on VSMC. We show the discriminatory power of SBT by unraveling with unprecedented molecular detail the coordinated mechanisms taking place during AngII stimulation of VSMC. The robustness of the model is supported by remarkable maintenance of coordination in the same categories in different VSMC preparations over time. Although from our data we cannot judge whether this regulation occurs at transcriptional or posttranscriptional levels, our finding that the vast majority of coordinated categories are protein complexes, metabolic pathways or proteins forming interaction networks, suggests that protein-protein interactions play an important role in producing tight coregulation of protein levels. This conclusion is supported by detection of proteins with a seemingly uncoordinated behavior at the earliest time points, whereas the coordinated response elicited by AngII in VSMC builds up gradually, suggesting an adjustment of the regulatory protein machinery.

Some of the proteins described here to be affected by AngII have been previously observed as AngII sensitive in smooth muscle and other type of cells, including calponin (112), contractile proteins (113), small GTP-binding proteins (114), Rho-associated protein kinase (115), MAP kinases (116–118), and histones (119–122). However, the vast majority of the proteins detected here as being implicated in the coordinated action of

AngII have not been described previously, and all of them together provide for the first time a global picture of the early phenotypic alterations taking place in VSMC. The increases in the protein synthesis, folding and turnover machineries suggest the induction of a hypertrophic phenotype where for the first time we show the ability of AngII to activate spliceosome and prefoldin subunits. This phenotype is consistent with the observed reinforcement in contractility and migration, where we describe a novel effect of AngII as activator of the 5HT3 type receptor mediated signaling pathway, the complex-V of electron transport chain and the proteins of Arp2/3 complex. The decrease in fibrillar collagen during early times of AngII activation is also a completely new finding in the sense that it is opposite to the upregulatory effect produced by AngII on VSMC collagens at longer incubation times (48 h) (123). Finally, we also detect a general repression of proliferative and secretory protein categories, including the DNA-dependent ATPase MCM complex, the glutamine amidotransferase, and the complement, gluconeogenesis and homocysteine biosynthesis pathways, which have never been observed before to be downregulated by AngII in VSMC. These results therefore suggest that at early times AngII induces migration and inhibits proliferation in VSMC. This characterization of novel signaling pathways and effector molecules regulated by AngII in VSMCs during the early phase response might facilitate the identification of therapeutic targets for intervention in many cardiovascular diseases that occur with vascular wall remodeling.

Acknowledgments—We thank S. Bartlett for English editing.

* This work was supported by grant BIO2012–37926 from the Spanish Ministry of Economy and Competitiveness and grant PRB2 (IPT13/0001 - ISCIII-SGEFI/FEDER, ProteoRed) and RD12/0042/0056, RD12/0042/0054, RD12/0042/0028, and RD12/0042/0022 from RIC-Red Temática de Investigación Cooperativa en Salud (RETICS), Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III, and co-funding by Fondo Europeo de Desarrollo Regional (FEDER). SM is supported by the Fundación La Marató TV3 (122532). MTH was supported by a fellowship from the Spanish Ministry of Economy and Competitiveness. The CNIC is supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the Pro CNIC Foundation, and is a Severo Ochoa Center of Excellence (MINECO award SEV-2015-0505).

[S] This article contains supplemental Figs. S1 to S7 and Tables S1 to S8.

§ To whom correspondence should be addressed: Cardiovascular Proteomics Laboratory; Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC); Melchor Fernández Almagro, 3; 28029 Madrid, Spain. Tel.: (+34) 914531200; E-mail: ebonzon@cnic.es, jvazquez@cnic.es.

¶ In alphabetical order; these authors contributed equally to this work.

Email addresses: Fernando García-Marqués fernando.garcia@externo.cnic.es; Marco Trevisan-Herraz mtrevisan@cnic.es; Sara Martínez-Martínez smartinez@cnic.es; Emilio Camafeita ecamafeita@cnic.es; Inmaculada Jorge Cerrudo inmaculada.jorge@cnic.es; Juan Antonio López jalopez@cnic.es; Nerea Méndez Barbero nmendez@cnic.es; Simón Méndez Ferrer simon.mendez-ferrer@cnic.es; Miguel Angel del Pozo madelpozo@cnic.es; Borja Ibáñez bibanez@cnic.es; Vicente Andrés García vandres@cnic.es; Francisco Sánchez-Madrid fsanchez-madrid@cnic.es; Juan Miguel Redondo jmredondo@cnic.es.

REFERENCES

1. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 370–377
2. Ihmels, J., Levy, R., and Barkai, N. (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nat. Biotechnol.* **22,** 86–92
3. Wei, H., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G. P., Somerville, C., and Loraine, A. (2006) Transcriptional coordination of the metabolic network in Arabidopsis. *Plant Physiol.* 762–774
4. Sprinzak, E., Cokus, S. J., Yeates, T. O., Eisenberg, D., and Pellegrini, M. (2009) Detecting coordinated regulation of multiprotein complexes using logic analysis of gene expression. *BMC Syst. Biol.* **3,** 115
5. Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. U.S.A.* 2981–2986
6. Gandhi, S. J., Zenklusen, D., Lionnet, T., and Singer, R. H. (2011) Transcription of functionally related constitutive genes is not coordinated. *Nat. Struct. Mol. Biol.* **18,** 27–34
7. Newman, J., Ghaemmaghami, S., and Ihmels, J. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*
8. Carmi, S., Levanon, E. Y., and Eisenberg, E. (2009) Efficiency of complex production in changing environment. *BMC Syst. Biol.* 3
9. Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006) Noise in protein expression scales with natural protein abundance. *Nat. Genet.* 636–643
10. Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S. (2003) Global analysis of protein expression in yeast. *Nature* 737–741
11. Carmi, S., Levanon, E. Y., Havlin, S., and Eisenberg, E. (2006) Connectivity and expression in protein networks: proteins in a complex are uniformly expressed. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 031909–031906
12. Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bähler, J. (2012) Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* 671–683
13. Foster, L. J., de Hoog, C. L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V. K., and Mann, M. (2006) A mammalian organelle map by protein correlation profiling. *Cell* 187–199
14. Hansson, J., Rafiee, M. R., Reiland, S., Polo, J. M., Gehring, J., Okawa, S., Huber, W., Hochedlinger, K., and Krijgsveld, J. (2012) Highly coordinated proteome dynamics during reprogramming of somatic cells to pluripotency. *Cell Rep.* 1579–1592
15. Wu, Y., Williams, E. G., Dubuis, S., Mottis, A., Jovaisaite, V., Houten, S. M., Argmann, C. A., Faridi, P., Wolski, W., Kutalik, Z., Zamboni, N., Auwerx, J., and Aebersold, R. (2014) Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell* 1415–1430
16. Lacolley, P., Regnault, V., Nicoletti, A., Li, Z., and Michel, J. B. (2012) The vascular smooth muscle cell in arterial pathology: a cell that can take on multiple roles. *Cardiovasc. Res.* **95,** 194–204
17. Touyz, R. M. (2005) Intracellular mechanisms involved in vascular remodeling of resistance arteries in hypertension: role of angiotensin II. *Exp. Physiol.* **90,** 449–455
18. Daugherty, A., Manning, M. W., and Cassis, L. A. (2000) Angiotensin II promotes atherosclerotic lesions and aneurysms in apolipoprotein E-deficient mice. *J. Clin. Invest.* **105,** 1605–1612
19. Heeneman, S., Sluimer, J. C., and Daemen, M. J. (2007) Angiotensin-converting enzyme and vascular remodeling. *Circ. Res.* **101,** 441–454
20. Weintraub, N. L. (2009) Understanding abdominal aortic aneurysm. *N. Engl. J. Med.* **361,** 1114–1116
21. Navarro, P., Trevisan-Herraz, M., Bonzon-Kulichenko, E., Nunez, E., Martinez-Acedo, P., Perez-Hernandez, D., Jorge, I., Mesa, R., Calvo, E., Carrascal, M., Hernaez, M. L., Garcia, F., Barcena, J. A., Ashman, K., Abian, J., Gil, C., Redondo, J. M., and Vazquez, J. (2014) General

statistical framework for quantitative proteomics by stable isotope labeling. *J. Proteome Res.* **13,** 1234–1247
22. Marquardt, D. W. (1963) An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.* **11,** 431–441
23. Anderson, T., Hankin, R., and Killworth, P. (2008) Beyond the Durfee square: enhancing the h-index to score total publication output. *Scientometrics* **76,** 577–588
24. Isern, J., Martin-Antonio, B., Ghazanfari, R., Martin, A. M., Lopez, J. A., del Toro, R., Sanchez-Aguilera, A., Arranz, L., Martin-Perez, D., Suarez-Lledo, M., Marin, P., Van Pel, M., Fibbe, W. E., Vazquez, J., Sheding, S., Urbano-Ispizua, A., and Mendez-Ferrer, S. (2013) Self-renewing human bone marrow mesenspheres promote hematopoietic stem cell expansion. *Cell Rep.* **3,** 1714–1724
25. Wieckowski, M. R., Giorgi, C., Lebiedzinska, M., Duszynski, J., and Pinton, P. (2009) Isolation of mitochondria-associated membranes and mitochondria from animal tissues and cells. *Nat. Protoc.* **4,** 1582–1590
26. Varela, I., Cadinanos, J., Pendas, A. M., Gutierrez-Fernandez, A., Folgueras, A. R., Sanchez, L. M., Zhou, Z., Rodriguez, F. J., Stewart, C. L., Vega, J. A., Tryggvason, K., Freije, J. M., and Lopez-Otin, C. (2005) Accelerated aging in mice deficient in Zmpste24 protease is linked to p53 signaling activation. *Nature* **437,** 564–568
27. Danielsen, M., Hornshoj, H., Siggers, R. H., Jensen, B. B., van Kessel, A. G., and Bendixen, E. (2007) Effects of bacterial colonization on the porcine intestinal proteome. *J. Proteome Res.* **6,** 2596–2604
28. Garcia-Prieto, J., Garcia-Ruiz, J. M., Sanz-Rosa, D., Pun, A., Garcia-Alvarez, A., Davidson, S. M., Fernandez-Friera, L., Nuno-Ayala, M., Fernandez-Jimenez, R., Bernal, J. A., Izquierdo-Garcia, J. L., Jimenez-Borreguero, J., Pizarro, G., Ruiz-Cabello, J., Macaya, C., Fuster, V., Yellon, D. M., and Ibanez, B. (2014) beta3 adrenergic receptor selective stimulation during ischemia/reperfusion improves cardiac function in translational models through inhibition of mPTP opening in cardiomyocytes. *Basic Res. Cardiol.* **109,** 422
29. Gonzalez-Granado, J. M., Silvestre-Roig, C., Rocha-Perugini, V., Trigueros-Motos, L., Cibrian, D., Morlino, G., Blanco-Berrocal, M., Osorio, F. G., Freije, J. M., Lopez-Otin, C., Sanchez-Madrid, F., and Andres, V. (2014) Nuclear envelope lamin-A couples actin dynamics with immunological synapse architecture and T cell activation. *Sci. Signal.* **7,** ra37
30. Ray, J. L., Leach, R., Herbert, J. M., and Benson, M. (2001) Isolation of vascular smooth muscle cells from a single murine aorta. *Methods Cell Sci.* **23,** 185–188
31. Bonzon-Kulichenko, E., Perez-Hernandez, D., Nunez, E., Martinez-Acedo, P., Navarro, P., Trevisan-Herraz, M., Ramos Mdel, C., Sierra, S., Martinez-Martinez, S., Ruiz-Meana, M., Miro-Casas, E., Garcia-Dorado, D., Redondo, J. M., Burgos, J. S., and Vazquez, J. (2011) A robust method for quantitative high-throughput analysis of proteomes by 18O labeling. *Mol. Cell. Proteomics* **10,** M110 003335
32. Leyfer, D., and Weng, Z. (2005) Genome-wide decoding of hierarchical modular structure of transcriptional regulation by *cis*-element and expression clustering. *Bioinformatics* **2,** ii197–203
33. Martinez-Bartolome, S., Navarro, P., Martin-Maroto, F., Lopez-Ferrer, D., Ramos-Fernandez, A., Villar, M., Garcia-Ruiz, J. P., and Vazquez, J. (2008) Properties of average score distributions of SEQUEST: the probability ratio method. *Mol. Cell. Proteomics* **7,** 1135–1145
34. Navarro, P., and Vazquez, J. (2009) A refined method to calculate false discovery rates for peptide identification using decoy databases. *J. Proteome Res.* **8,** 1792–1796
35. Bonzon-Kulichenko, E., Garcia-Marques, F., Trevisan-Herraz, M., and Vazquez, J. (2015) Revisiting peptide identification by high-accuracy mass spectrometry: problems associated with the use of narrow mass precursor windows. *J. Proteome Res.* **14,** 700–710
36. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* **473,** 337–342
37. Ficenec, D., Osborne, M., Pradines, J., Richards, D., Felciano, R., Cho, R. J., Chen, R. O., Liefeld, T., Owen, J., Ruttenberg, A., Reich, C., Horvath, J., and Clark, T. (2003) Computational knowledge integration in biopharmaceutical research. *Brief. Bioinformatics* **4,** 260–278
38. Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., Lowry, S. F., Inflamm, and Host Response to Injury Large Scale Collab. Res, P. (2005) A network-based analysis of systemic inflammation in humans. *Nature* **437,** 1032–1037
39. Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O. N., Stumpflen, V., and Mewes, H. W. (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* **36,** D646–650
40. Huang da, W., Sherman, B. T., Zheng, X., Yang, J., Imamichi, T., Stephens, R., and Lempicki, R. A. (2009) Extracting biological meaning from large gene lists with DAVID. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis. [et al.]* Chapter 13, Unit 13 11
41. Vizcaino, J. A., Cote, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41,** D1063–1069
42. Reinheckel, T., Sitte, N., Ullrich, O., Kuckelkorn, U., Davies, K. J., and Grune, T. (1998) Comparative resistance of the 20S and 26S proteasome to oxidative stress. *Biochem. J.* **335,** 637–642
43. Thorpe, G. W., Fong, C. S., Alic, N., Higgins, V. J., and Dawes, I. W. (2004) Cells have distinct mechanisms to maintain protection against different reactive oxygen species: oxidative-stress-response genes. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 6564–6569
44. Lapinskas, P. J., Cunningham, K. W., Liu, X. F., Fink, G. R., and Culotta, V. C. (1995) Mutations in PMR1 suppress oxidative damage in yeast cells lacking superoxide dismutase. *Mol. Cell. Biol.* **15,** 1382–1388
45. Liu, X. F., and Culotta, V. C. (1994) The requirement for yeast superoxide dismutase is bypassed through mutations in BSD2, a novel metal homeostasis gene. *Mol. Cell. Biol.* **14,** 7037–7045
46. Shenton, D., Smirnova, J. B., Selley, J. N., Carroll, K., Hubbard, S. J., Pavitt, G. D., Ashe, M. P., and Grant, C. M. (2006) Global translational responses to oxidative stress impact upon multiple levels of protein synthesis. *J. Biol. Chem.* **281,** 29011–29021
47. Daemen, M. J., Lombardi, D. M., Bosman, F. T., and Schwartz, S. M. (1991) Angiotensin II induces smooth muscle cell proliferation in the normal and injured rat arterial wall. *Circ. Res.* **68,** 450–456
48. Dubey, R. K., Jackson, E. K., and Luscher, T. F. (1995) Nitric oxide inhibits angiotensin II-induced migration of rat aortic smooth muscle cell. Role of cyclic-nucleotides and angiotensin1 receptors. *J. Clin. Invest.* **96,** 141–149
49. Sprinzak, E., Cokus, S. J., Yeates, T. O., Eisenberg, D., and Pellegrini, M. (2009) Detecting coordinated regulation of multiprotein complexes using logic analysis of gene expression. *BMC Syst. Biol.* **3,** 115
50. Simonis, N., Gonze, D., Orsi, C., van Helden, J., and Wodak, S. J. (2006) Modularity of the transcriptional response of protein complexes in yeast. *J. Mol. Biol.* **363,** 589–610
51. Kouzarides, T. (2002) Histone methylation in transcriptional control. *Curr. Opin. Genet. Dev.* **12,** 198–209
52. Chen, Y. I., Moore, R. E., Ge, H. Y., Young, M. K., Lee, T. D., and Stevens, S. W. (2007) Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors. *Nucleic Acids Res.* **35,** 3928–3944
53. Nishida, W., Kitami, Y., and Hiwada, K. (1993) cDNA cloning and mRNA expression of calponin and SM22 in rat aorta smooth muscle cells. *Gene* **130,** 297–302
54. Fu, H., Subramanian, R. R., and Masters, S. C. (2000) 14–3-3 proteins: structure, function, and regulation. *Annu. Rev. Pharmacol. Toxicol.* **40,** 617–647
55. McKay, M. M., Ritt, D. A., and Morrison, D. K. (2009) Signaling dynamics of the KSR1 scaffold complex. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 11022–11027
56. Maricq, A. V., Peterson, A. S., Brake, A. J., Myers, R. M., and Julius, D. (1991) Primary structure and functional expression of the 5HT3 receptor, a serotonin-gated ion channel. *Science* **254,** 432–437
57. Koves, T. R., Li, P., An, J., Akimoto, T., Slentz, D., Ilkayeva, O., Dohm, G. L., Yan, Z., Newgard, C. B., and Muoio, D. M. (2005) Peroxisome proliferator-activated receptor-gamma coactivator 1alpha-mediated metabolic remodeling of skeletal myocytes mimics exercise training and reverses lipid-induced mitochondrial inefficiency. *J. Biol. Chem.* **280,** 33588–33598
58. Somlyo, A. P., and Somlyo, A. V. (2003) Ca2+ sensitivity of smooth

muscle and nonmuscle myosin II: modulated by G proteins, kinases, and myosin phosphatase. *Physiolog. Rev.* **83,** 1325–1358
59. Takai, Y., Sasaki, T., and Matozaki, T. (2001) Small GTP-binding proteins. *Physiolog. Rev.* **81,** 153–208
60. Ohtsu, H., Mifune, M., Frank, G. D., Saito, S., Inagami, T., Kim-Mitsuyama, S., Takuwa, Y., Sasaki, T., Rothstein, J. D., Suzuki, H., Nakashima, H., Woolfolk, E. A., Motley, E. D., and Eguchi, S. (2005) Signal-crosstalk between Rho/ROCK and c-Jun NH2-terminal kinase mediates migration of vascular smooth muscle cells stimulated by angiotensin II. *Arterioscler. Thromb. Vasc. Biol.* **25,** 1831–1836
61. Frank, S. R., Hatfield, J. C., and Casanova, J. E. (1998) Remodeling of the actin cytoskeleton is coordinately regulated by protein kinase C and the ADP-ribosylation factor nucleotide exchange factor ARNO. *Mol. Biol. Cell* **9,** 3133–3146
62. Franco, M., Peters, P. J., Boretto, J., van Donselaar, E., Neri, A., D'Souza-Schorey, C., and Chavrier, P. (1999) EFA6, a sec7 domain-containing exchange factor for ARF6, coordinates membrane recycling and actin cytoskeleton organization. *EMBO J.* **18,** 1480–1491
63. Boshans, R. L., Szanto, S., van Aelst, L., and D'Souza-Schorey, C. (2000) ADP-ribosylation factor 6 regulates actin cytoskeleton remodeling in coordination with Rac1 and RhoA. *Mol. Cell. Biol.* 20, 3685–3694
64. Nishiya, N., Kiosses, W. B., Han, J., and Ginsberg, M. H. (2005) An alpha4 integrin-paxillin-Arf-GAP complex restricts Rac activation to the leading edge of migrating cells. *Nat. Cell Biol.* **7,** 343–352
65. Santy, L. C., and Casanova, J. E. (2001) Activation of ARF6 by ARNO stimulates epithelial cell migration through downstream activation of both Rac1 and phospholipase D. *J. Cell Biol.* **154,** 599–610
66. Gimona, M., Kaverina, I., Resch, G. P., Vignal, E., and Burgstaller, G. (2003) Calponin repeats regulate actin filament stability and formation of podosomes in smooth muscle cells. *Mol. Biol. Cell* **14,** 2482–2491
67. Fluck, M., Mund, S. I., Schittny, J. C., Klossner, S., Durieux, A. C., and Giraud, M. N. (2008) Mechano-regulated tenascin-C orchestrates muscle repair. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 13662–13667
68. Yabkowitz, R., Mansfield, P. J., Ryan, U. S., and Suchard, S. J. (1993) Thrombospondin mediates migration and potentiates platelet-derived growth factor-dependent migration of calf pulmonary artery smooth muscle cells. *J. Cell. Physiol.* **157,** 24–32
69. Tanaka, S., Koyama, H., Ichii, T., Shioi, A., Hosoi, M., Raines, E. W., and Nishizawa, Y. (2002) Fibrillar collagen regulation of plasminogen activator inhibitor-1 is involved in altered smooth muscle cell migration. *Arterioscler. Thromb. Vasc. Biol.* **22,** 1573–1578
70. Zhang, J., Zou, F., Tang, J., Zhang, Q., Gong, Y., Wang, Q., Shen, Y., Xiong, L., Breyer, R. M., Lazarus, M., Funk, C. D., and Yu, Y. (2013) Cyclooxygenase-2-derived prostaglandin E(2) promotes injury-induced vascular neointimal hyperplasia through the E-prostanoid 3 receptor. *Circ. Res.* **113,** 104–114
71. Coxon, A., Maundrell, K., and Kearsey, S. E. (1992) Fission yeast cdc21+ belongs to a family of proteins involved in an early step of chromosome replication. *Nucleic Acids Res.* **20,** 5571–5577
72. Hu, B., Burkhart, R., Schulte, D., Musahl, C., and Knippers, R. (1993) The P1 family: a new class of nuclear mammalian proteins related to the yeast Mcm replication proteins. *Nucleic Acids Res.* **21,** 5289–5293
73. Lacey, J. M., and Wilmore, D. W. (1990) Is glutamine a conditionally essential amino acid? *Nutr. Rev.* **48,** 297–309
74. Chiang, J. K., Sung, M. L., Yu, H. R., Chang, H. I., Kuo, H. C., Tsai, T. C., Yen, C. K., and Chen, C. N. (2011) Homocysteine induces smooth muscle cell proliferation through differential regulation of cyclins A and D1 expression. *J. Cell. Physiol.* **226,** 1017–1026
75. Chiang, P. K., Gordon, R. K., Tal, J., Zeng, G. C., Doctor, B. P., Pardhasaradhi, K., and McCann, P. P. (1996) S-Adenosylmethionine and methylation. *FASEB J.* **10,** 471–480
76. Lindsay, H., and Adams, R. L. (1996) Spreading of methylation along DNA. *Biochem. J.* **320,** 473–478
77. Strahl, B. D., Grant, P. A., Briggs, S. D., Sun, Z. W., Bone, J. R., Caldwell, J. A., Mollah, S., Cook, R. G., Shabanowitz, J., Hunt, D. F., and Allis, C. D. (2002) Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. *Mol. Cell. Biol.* **22,** 1298–1306
78. Davis, C. D., and Ross, S. A. (2007) Dietary components impact histone modifications and cancer risk. *Nutr. Rev.* **65,** 88–94
79. Rice, J. C., Briggs, S. D., Ueberheide, B., Barber, C. M., Shabanowitz, J., Hunt, D. F., Shinkai, Y., and Allis, C. D. (2003) Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains. *Mol. Cell* **12,** 1591–1598
80. Gasque, P., Neal, J. W., Singhrao, S. K., McGreal, E. P., Dean, Y. D., Van, B. J., and Morgan, B. P. (2002) Roles of the complement system in human neurodegenerative disorders: pro-inflammatory and tissue remodeling activities. *Mol. Neurobiol.* **25,** 1–17
81. Lin, Z. H., Fukuda, N., Jin, X. Q., Yao, E. H., Ueno, T., Endo, M., Saito, S., Matsumoto, K., and Mugishima, H. (2004) Complement 3 is involved in the synthetic phenotype and exaggerated growth of vascular smooth muscle cells from spontaneously hypertensive rats. *Hypertension* **44,** 42–47
82. Paul, R. J. (1989) Smooth muscle energetics. *Annu. Rev. Physiol.* **51,** 331–349
83. Dzau, V. J. (1998) Mechanism of protective effects of ACE inhibition on coronary artery disease. *Eur. Heart J.* **19,** J2–6
84. Gao, P., Qian, D. H., Li, W., and Huang, L. (2009) NPRA-mediated suppression of AngII-induced ROS production contribute to the antiproliferative effects of B-type natriuretic peptide in VSMC. *Mol. Cell. Biochem.* **324,** 165–172
85. Yang, H., Ara, A. I., Magilnick, N., Xia, M., Ramani, K., Chen, H., Lee, T. D., Mato, J. M., and Lu, S. C. (2008) Expression pattern, regulation, and functions of methionine adenosyltransferase 2beta splicing variants in hepatoma cells. *Gastroenterology* **134,** 281–291
86. Zahradka, P., Werner, J. P., Buhay, S., Litchie, B., Helwer, G., and Thomas, S. (2002) NF-kappaB activation is essential for angiotensin II-dependent proliferation and migration of vascular smooth muscle cells. *J. Mol. Cell. Cardiol.* **34,** 1609–1621
87. Dai, G., Vaughn, S., Zhang, Y., Wang, E. T., Garcia-Cardena, G., and Gimbrone, M. A., Jr. (2007) Biomechanical forces in atherosclerosis-resistant vascular regions regulate endothelial redox balance via phosphoinositol 3-kinase/Akt-dependent activation of Nrf2. *Circ. Res.* **101,** 723–733
88. Chen, X. L., Varner, S. E., Rao, A. S., Grey, J. Y., Thomas, S., Cook, C. K., Wasserman, M. A., Medford, R. M., Jaiswal, A. K., and Kunsch, C. (2003) Laminar flow induction of antioxidant response element-mediated genes in endothelial cells. A novel anti-inflammatory mechanism. *J. Biol. Chem.* **278,** 703–711
89. Kim, Y. J., Ahn, J. Y., Liang, P., Ip, C., Zhang, Y., and Park, Y. M. (2007) Human prx1 gene is a target of Nrf2 and is upregulated by hypoxia/reoxygenation: implication to tumor biology. *Cancer Res.* **67,** 546–554
90. Volanakis, J. E. (1995) Transcriptional regulation of complement genes. *Annu. Rev. Immunol.* **13,** 277–305
91. Jitrapakdee, S. (2012) Transcription factors and coactivators controlling nutrient and hormonal regulation of hepatic gluconeogenesis. *Int. J. Biochem. Cell Biol.* **44,** 33–45
92. Sharifi, B. G., LaFleur, D. W., Pirola, C. J., Forrester, J. S., and Fagin, J. A. (1992) Angiotensin II regulates tenascin gene expression in vascular smooth muscle cells. *J. Biol. Chem.* **267,** 23910–23915
93. Majack, R. A., Mildbrandt, J., and Dixit, V. M. (1987) Induction of thrombospondin messenger RNA levels occurs as an immediate primary response to platelet-derived growth factor. *J. Biol. Chem.* **262,** 8821–8825
94. Scott-Burden, T., Resink, T. J., Hahn, A. W., and Buhler, F. R. (1990) Induction of thrombospondin expression in vascular smooth muscle cells by angiotensin II. *J. Cardiovasc. Pharmacol.* **7,** S17–20
95. Mouneimne, G., Hansen, S. D., Selfors, L. M., Petrak, L., Hickey, M. M., Gallegos, L. L., Simpson, K. J., Lim, J., Gertler, F. B., Hartwig, J. H., Mullins, R. D., and Brugge, J. S. (2012) Differential remodeling of actin cytoskeleton architecture by profilin isoforms leads to distinct effects on cell migration and invasion. *Cancer Cell* **22,** 615–630
96. Tokuo, H., and Ikebe, M. (2004) Myosin X transports Mena/VASP to the tip of filopodia. *Biochem. Biophys. Res. Commun.* **319,** 214–220
97. Zhang, H., Berg, J. S., Li, Z., Wang, Y., Lang, P., Sousa, A. D., Bhaskar, A., Cheney, R. E., and Stromblad, S. (2004) Myosin-X provides a motor-based link between integrins and the cytoskeleton. *Nat. Cell Biol.* **6,** 523–531
98. Kerber, M. L., Jacobs, D. T., Campagnola, L., Dunn, B. D., Yin, T., Sousa, A. D., Quintero, O. A., and Cheney, R. E. (2009) A novel form of motility in filopodia revealed by imaging myosin-X at the single-molecule level. *Curr. Biol.* **19,** 967–973

99. Xia, C., Bao, Z., Tabassam, F., Ma, W., Qiu, M., Hua, S., and Liu, M. (2000) GCIP, a novel human grap2 and cyclin D interacting protein, regulates E2F-mediated transcriptional activity. *J. Biol. Chem.* **275,** 20942–20948
100. Chang, T. W., Chen, C. C., Chen, K. Y., Su, J. H., Chang, J. H., and Chang, M. C. (2008) Ribosomal phosphoprotein P0 interacts with GCIP and overexpression of P0 is associated with cellular proliferation in breast and liver carcinoma cells. *Oncogene* **27,** 332–338
101. Lee, K. M., Seo, H. Y., Kim, M. K., Min, A. K., Ryu, S. Y., Kim, Y. N., Park, Y. J., Choi, H. S., Lee, K. U., Park, W. J., Park, K. G., and Lee, I. K. (2010) Orphan nuclear receptor small heterodimer partner inhibits angiotensin II- stimulated PAI-1 expression in vascular smooth muscle cells. *Exp. Mol. Med.* **42,** 21–29
102. Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10,** 48
103. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–452
104. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550
105. Khatri, P., Sirota, M., and Butte, A. J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* e1002375
106. Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids Res.* **37,** 1–13
107. Khatri, P., and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21,** 3587–3595
108. Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34,** 267–273
109. Barry, W. T., Nobel, A. B., and Wright, F. A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21,** 1943–1949
110. Jiang, Z., and Gentleman, R. (2007) Extensions to gene set enrichment. *Bioinformatics* **23,** 306–313
111. Jorge, I., Navarro, P., Martinez-Acedo, P., Nunez, E., Serrano, H., Alfranca, A., Redondo, J. M., and Vazquez, J. (2009) Statistical model to analyze quantitative proteomics data obtained by 18O/16O labeling and linear ion trap mass spectrometry: application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells. *Mol. Cell. Proteomics* **8,** 1130–1149
112. di Gioia, C. R., van de Greef, W. M., Sperti, G., Castoldi, G., Todaro, N., Ierardi, C., Pieruzzi, F., and Stella, A. (2000) Angiotensin II increases calponin expression in cultured rat vascular smooth muscle cells. *Biochem. Biophys. Res. Commun.* **279,** 965–969
113. Turla, M. B., Thompson, M. M., Corjay, M. H., and Owens, G. K. (1991) Mechanisms of angiotensin II- and arginine vasopressin-induced increases in protein synthesis and content in cultured rat aortic smooth muscle cells. Evidence for selective increases in smooth muscle isoactin expression. *Circ. Res.* **68,** 288–299
114. Ohtsu, H., Suzuki, H., Nakashima, H., Dhobale, S., Frank, G. D., Motley, E. D., and Eguchi, S. (2006) Angiotensin II signal transduction through small GTP-binding proteins: mechanism and significance in vascular smooth muscle cells. *Hypertension* **48,** 534–540
115. Rattan, S., Puri, R. N., and Fan, Y. P. (2003) Involvement of rho and rho-associated kinase in sphincteric smooth muscle contraction by angiotensin II. *Exp. Biol. Med.* **228,** 972–981
116. Duff, J. L., Berk, B. C., and Corson, M. A. (1992) Angiotensin II stimulates the pp44 and pp42 mitogen-activated protein kinases in cultured rat aortic smooth muscle cells. *Biochem. Biophys. Res. Commun.* **188,** 257–264
117. Ishida, Y., Kawahara, Y., Tsuda, T., Koide, M., and Yokoyama, M. (1992) Involvement of MAP kinase activators in angiotensin II-induced activation of MAP kinases in cultured vascular smooth muscle cells. *FEBS Letters* **310,** 41–45
118. Sadoshima, J., Qiu, Z., Morgan, J. P., and Izumo, S. (1995) Angiotensin II and other hypertrophic stimuli mediated by G protein-coupled receptors activate tyrosine kinase, mitogen-activated protein kinase, and 90-kD S6 kinase in cardiac myocytes. The critical role of Ca(2+)-dependent signaling. *Circ. Res.* **76,** 1–15
119. Fukuda, K., and Izumo, S. (1998) Angiotensin II potentiates DNA synthesis in AT-1 transformed cardiomyocytes. *J. Mol. Cell. Cardiol.* **30,** 2069–2080
120. Elliott, M. E. (1990) Phosphorylation of adrenal histone H3 is affected by angiotensin, ACTH, dibutyryl cAMP, and atrial natriuretic peptide. *Life Sci.* **46,** 1479–1488
121. Xu, X., Ha, C. H., Wong, C., Wang, W., Hausser, A., Pfizenmaier, K., Olson, E. N., McKinsey, T. A., and Jin, Z. G. (2007) Angiotensin II stimulates protein kinase D-dependent histone deacetylase 5 phosphorylation and nuclear export leading to vascular smooth muscle cell hypertrophy. *Arterioscler. Thromb. Vasc. Biol.* **27,** 2355–2362
122. Chu, C. H., Lo, J. F., Hu, W. S., Lu, R. B., Chang, M. H., Tsai, F. J., Tsai, C. H., Weng, Y. S., Tzang, B. S., and Huang, C. Y. (2012) Histone acetylation is essential for ANG-II-induced IGF-IIR gene expression in H9c2 cardiomyoblast cells and pathologically hypertensive rat heart. *J. Cell. Physiol.* **227,** 259–268
123. Wang, W., Huang, X. R., Canlas, E., Oka, K., Truong, L. D., Deng, C., Bhowmick, N. A., Ju, W., Bottinger, E. P., and Lan, H. Y. (2006) Essential role of Smad3 in angiotensin II-induced vascular fibrosis. *Circ. Res.* **98,** 1032–1039
124. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389,** 251–260
125. Zaric, B., Chami, M., Remigy, H., Engel, A., Ballmer-Hofer, K., Winkler, F. K., and Kambach, C. (2005) Reconstitution of two recombinant LSm protein complexes reveals aspects of their architecture, assembly, and function. *J. Biol. Chem.* **280,** 16066–16075
126. Pillai, R. S., Grimmler, M., Meister, G., Will, C. L., Luhrmann, R., Fischer, U., and Schumperli, D. (2003) Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing. *Genes Dev.* **17,** 2321–2333

4. Brief account of other results, published or unpublished, concerning this work

4.1 Summary

The three papers presented in this chapter can be seen together as a great step in the bioinformatics tools applied to proteomics, and were key in the evolution of the techniques and procedures used in the laboratory where they have been implemented for the last four years. However, the results and applications are not limited to those here exposed.

There are two additional papers (see Appendix 1, subsections 1.1 and 1.2) to which most or part of the research presented here has been applied, providing feedback for the development of the models, and two more papers in preparation to present the software developed to implement the algorithms and make them readily available to other users. In addition, the computational resources resulting from this work have been used on a day-to-day basis in the Cardiovascular Laboratory at the Centro Nacional de Investigaciones Cardiovasculares (CNIC), being one of the main tools for the research performed. Furthermore, the software developed is available and has been shared with any research group requesting for it, and will be released to the public on the internet as soon as the mentioned papers in preparation are published; in this sense, the use of the statistical models and the software extends to research in other laboratories and areas (Ezkurdia, 2015; Latorre-Pellicer, 2016; Mateos-Hernández, 2016), and is not limited to the applications presented in this work.

4.2 The human HDL proteome displays high inter-individual variability and is altered dynamically in response to angioplasty-induced atheroma plaque rupture

The WSPP model and the GIA were used to apply the Fundamental Workflow and integrate the experiments from the different individuals. Additionally, I used the feedback to improve and modify the associated software in order to adapt it to the needs of the research performed. The SBT was used as well to perform the systems biology analysis.

The study of the proteome associated to high-density lipoproteins (HDL), comparing samples from different patients of coronary disease before and after an angioplasty operation, reinforced the idea that HDLs are dynamic platforms

interchanging different molecules, facilitating the search for cardiovascular biomarkers. (Jorge, 2014)

4.3 Quantitative HDL proteomics identifies peroxiredoxin-6 as a biomarker of human abdominal aortic aneurysm

The WSPP statistical model and the Fundamental Workflow were used in this work to perform the quantitative proteomics experiments and the bioinformatic analysis. The GIA was also essential to integrate data from different experiments, so they could be analysed and compared. The SBT was also used to perform the systems biology.

The study of the proteome associated to high-density lipoproteins (HDL) revealed that the mechanism behind abdominal aortic aneurysm is based on a redox imbalance, describing for the first time the PRDX6 antioxidant protein as a systemic biomarker for this disease. (Burillo, 2016)

4.4 QuiXoT: quantification and statistics of high-throughput proteomics by stable isotope labelling *(in preparation)*

In this paper the software package QuiXoT is presented. It has been extensively tested and used in our laboratory for the analysis of quantitative proteomics experiments, and is the first software that used the WSPP model. It covers the need for a software applying that statistical model, including convenient data visualisation tools to check the correct data modelling. I am the main collaborator in the second phase of development of the software presented (whose initiation and development in the first phase corresponds to Pedro Navarro), and the main author of the article. (*In preparation*)

4.5 SanXoT: a software package to allow the creation of limitless workflows in quantitative proteomics *(in preparation)*

The goal of this paper will be presenting the SanXoT software platform, which implements an advanced version of the WSPP model. Compared to the QuiXoT software package, the main advantage of SanXoT is its flexible design, as it is presented as a collection of small independent software tools performing simple operations, working together using simple text files as input/output. The main program, also called SanXoT, is dedicated to the application of the GIA, allowing unlimited possibilities for the design of workflows in quantitative proteomics. Among these possibilities are the SBT and the experiment merging workflows, whose practical use is possible thanks to this software (intending by *experiment merging* the combination of

data from different experiments in such a way that the final data are independent of the experiment, calculating the inter-experiment variance and outliers across experiments). Additionally, specific software for systems biology and experiment merging is included in the package. More about this software can be read in section 4 of the Discussion chapter.

At the moment, I am the sole developer of this software, and its development has advanced in parallel to its extensive use and thanks to the feedback from the day-to-day use by all colleagues of the laboratory. (*In preparation*)

Discussion

It was effortless. It was easy to play with these things. It was like uncorking a bottle: everything flowed out effortlessly

Richard Feynman, «*Surely you're joking, Mr. Feynman*»

1. Advances in protein identification by MS/MS using high-accuracy precursor mass information

1.1 A problem worth careful consideration

As it has been stated in the introduction (subsection 1.2), the algorithm to validate the peptide identification and calculate the FDR was in need of revision. With the advent of high-resolution mass spectrometry, accuracy and speed of instruments improved at a high pace in recent years, exposing deficiencies in the algorithms that had not been observable previously. Suddenly, the amount and quality of spectra increased, which enabled the identification of many more peptides, although paradoxically the number of *good* peptide identifications decayed. In this paper we proposed a method to improve the quality of peptide-spectrum matches using current algorithms.

1.2 Independent-scores vs database-dependent scores

We focussed on a representative population of 10,000 spectra and analysed carefully the effect of search space on the accuracy with which the number of false PSMs is estimated using decoy databases. When we used independent scores, i.e. scores that only depend on the spectrum and the sequence that is matched, we were unable to detect a detrimental effect on this estimate even when searching against an average of only 200 sequence candidates. Indeed, estimates were so reproducible that we were able to make remarkably accurate predictions of FDR in diverse search conditions. However, we also found a marked difference in the behaviour of database-

dependent scores, defined here as the scores that take additional information from other sequence candidates. We observed that the peptide sequences identified using database-dependent scores when precursor mass tolerances are set in the low-ppm region can differ significantly from those obtained at wider tolerances, revealing an underlying inaccuracy in the estimation of FDR, which was due to a loss of reliability of database-dependent scores when the number of sequence candidates decreases. This basic principle was demonstrated using two classic, well-known and still widely-used searching engines, SEQUEST and Mascot, given that they both use independent and database-dependent scores. However, our results may be easily extrapolated to other searching algorithms (Nesvizhskii, 2010), once their scores are classified as being database-dependent or independent. A relevant example is Andromeda (Cox, 2011), a popular search engine that uses a database-independent score similar to Mascot's ions score and that, according to our line of reasoning, is not expected to lose reliability when very low precursor mass tolerances are used.

We should note here that, although increasing the mass precursor window is expected to produce more reliable results with database-dependent scores, it comes at the expense of considerable increases in search times. This is not only due to the increase in the number of sequence candidates that have to be scored, but also because binning candidates by their mass—a common method to accelerate peptide search—will be less computationally effective. This factor should be taken into account for a proper choice between database-dependent and independent scores.

1.3 Conclusion: database-dependent scores should be obtained using wide precursor mass windows, followed by a post-scoring precursor mass filtering

The results presented here complement, from a different line of evidence, previous reports that detected a decrease in the discriminative power of delta scores, reflecting increased variability due to a significant reduction in the number of candidate peptide sequences (Ding, 2008; Nesvizhskii, 2010). Based on the data obtained in this work, we suggest that when database-dependent scores are used, the high mass accuracy of modern mass spectrometres should not be exploited by filtering aggressively based on the precursor mass and then scoring peptide matches. Rather, and in line with the approaches used by an increasing number of authors (Beausoleil, 2006; Brosch, 2008; Ding, 2008; Hsieh, 2010; Nesvizhskii, 2010), we propose that more reliable results are to be expected when peptide matches are scored using wide precursor mass windows, avoiding inaccuracies derived from reduction of the search space, and then post-filtered according to their precursor mass.

2. The WSPP: a general statistical framework for the analysis of quantitative proteomics results.

2.1 Motivation to develop a statistical model

The second part of this work consists of the development of the WSPP statistical model (initialism for *Weighted Spectrum, Peptide and Protein*), the model employed in our laboratory for all quantitative proteomics experiments using stable isotope labelling (SIL). Having a suitable statistical model is not a secondary problem in high-throughput quantitative proteomics, it is a critical step in order to analyse experiments guaranteeing reproducibility and avoiding the presence of false positives when evaluating changes of expression. The nature of experiments in this area is such that using an unsuitable statistical model leads easily to non-normal distributions, with errors impossible to control or interpret, in many cases producing outliers in null-hypothesis experiments which call the whole process into question. Several statistical models had been described previous to the initial development of our model, but the practical use demonstrated that there was still scope for improvement.

2.2 General description of the WSPP statistical model

The model here presented was initially developed together with Pedro Navarro and published in 2014 (Navarro, 2014). The work presented here represents the continuation and culmination of the effort he started during the last stage of his PhD Thesis (whose second part consisted in developing a statistical model for ^{18}O labelling, further working on it to suit other instruments and SIL techniques). As we have demonstrated, it describes very accurately the technical variability of data for a representative set that includes the most common SIL methods. In addition, the model allows a systematic comparison and integration of data from different experiments. The general validity of the model was demonstrated by confronting 48 experimental distributions against 18 different null hypotheses. This statistical framework efficiently resolves specific issues: i) solves the problems of variance heterogeneity, data integration and undersampling, ii) provides a statistically-sound method for testing the quality of quantitative experiments while detecting experimental deviations, and iii) represents the first comprehensive standard proposed to date in the field of quantitative proteomics. Moreover, our results reinforce the idea that quantitative SIL experiments, if properly performed and analysed, are highly reproducible, irrespective of the labelling technique or MS platform used.

2.3 Analysis of the variance at each level: spectrum, peptide and protein

At the spectral level, the model analyses variance using a strategy different from variance-stabilisation normalisation procedures used to treat microarray (Huber, 2002) and iTRAQ data (Karp, 2010; Arntzen, 2011). Instead of transforming the data, the WSPP uses a two-parametre function to model the behaviour of variance, and from this function it directly assigns a variance to each one of the quantifications, keeping the original readings. This two-parametre modelling of variance is similar to that followed in a previous work to treat iTRAQ data (Zhang, 2010), although in that work the final analysis was made at the spectral level and no integration of the data at the peptide or protein levels was performed. The general applicability of our model to any SIL and MS combination confirms in the most general case a fundamental property of MS-based peptide-centric quantifications: the error produced during the quantitative SIL analysis of peptides generating the same intensity at a given MS detector, irrespective of their sequence or molecular structure, is constant and normally distributed. We believe that this property is of paramount importance in the field and simplifies interpretation of quantitative data produced by MS. Finally, a model that analyses specifically the variance at the spectral level has the advantage that it allows to control separately the error produced during MS analysis and quantification from that produced at the time of peptide or protein preparation, and thus may be used to check the proper functioning and calibration of the MS machines and even to detect chromatographic shifts due to incomplete coelution of peptide pairs. For instance, some of the MS conditions used in this work, such as the collision energy in PQD fragmentation, were optimised by selecting the ones that produced the minimum variance at the spectral level (Pedro Navarro, PhD Thesis).

Our results also suggest that, at least using the protocol followed here, the error produced during peptide preparation can be considered constant and normally-distributed, reinforcing previous results we obtained using other biological systems (Bonzon-Kulichenko, 2011b) and peptide preparation methods (Jorge, 2009). In the general case, the null hypothesis formulated here provides the basis for testing the validity of this assumption for any peptide preparation method. Analysis of variance at the peptide level may be very useful in practice to detect deviations from the expected behaviour during peptide preparation, such as those produced by artefacts like partial digestion or methionine oxidation (Bonzon-Kulichenko, 2011b). But it may also be used to assign a statistical significance level to abundance changes in postranslationally-modified peptides, such as Cys sites subjected to oxidative modification, as is demonstrated in another work (Martinez-Acedo, 2012).

Finally, the error at the protein level has been shown to follow the same trend in this and other biological systems (Bonzon-Kulichenko, 2011b), making it a very convenient starting point from which to analyse other error sources, such as biological or individual variance, which are highly dependent on experimental design and must be modelled separately in each case. Besides, since—at the protein level—the variances estimated in all the SIL experiments performed in this work (Table I) are very similar to those calculated in previous studies using ^{18}O labelling in several biological models (Jorge, 2009; Bonzon-Kulichenko, 2011b), including tissue extracts, and in general are below 0.01, this value may be used as a reference to determine whether the protein variability in a given preparation is higher than that expected for a conventional protein manipulation method. Thus, an increase in protein variance above the reference value not accompanied by a concomitant increase in peptide and spectrum variances indicates an increased heterogeneity in protein composition, not related to peptide manipulation or MS quantification. This heterogeneity may indicate technical problems related, for instance, to a protein preparation protocol involving too many steps, but may also reflect a high biological variability between the samples that are compared, as we have found in a previous work (Bonzon-Kulichenko, 2011a) and also when comparing human samples extracted from different individuals (Jorge, 2014). Note that the majority of existing models to analyse SIL data integrate the data into protein averages, without taking into account the variance at the lower levels, and then analyse as a whole the distribution of protein quantifications (Lin, 2006; Karp, 2010; Arntzen, 2011); in doing so, all the technical error sources are comprised in only one random variable and it is not possible to interpret results in terms of the variance at the protein level.

2.4 A framework to integrate quantitative information in hierarchical levels

One of the most important characteristics of the WSPP model is that it provides a general framework to integrate the quantitative information from one level to a superior level. When several spectra are integrated into a peptide average, the model takes into account the variance associated to each one of the spectra according to error propagation theory, so that the most accurate have a more significant contribution. This procedure simplifies the interpretation of data, since quantifications of poor quality have a negligible effect on the peptide average and do not need to be eliminated from the analysis. The variance assigned to each one of the peptides takes into account the variance at the spectrum level, which is diminished due to averaging of several spectra, and also the intrinsic variance associated to the process of peptide preparation. The

same is done when peptides are integrated into proteins; while peptide averaging diminishes the variance carried out from the spectrum and peptide levels, the protein average includes the variance produced by protein manipulation. In this sense, the constant variances at the spectrum, peptide and protein levels can be conceived as asymptotic errors, since they reflect the lowest error that can be achieved at each one of the levels. This concept of data integration is of general validity and can be extended to higher levels, where the effect of averaging is reflected when protein readings from different experiments are considered together. The decrease in variance of the data integrated at higher levels increase the statistical power to detect deviations from the null hypothesis. This explains why only a dozen of significant protein abundance changes are detected when the experiments were considered separately, while the alteration of more than one hundred proteins becomes evident when the data are integrated at higher levels (Figure 3 of second publication, page 49).

The use of weighted averages to calculate protein ratios is not new and reflects the current view that not all quantitative measurements have the same accuracy (Shadforth, 2005; Lin, 2006; Bantscheff, 2008; Cox and Mann, 2008; Karp, 2010; Zhang, 2010). However, the weighting scheme followed by our model is different from other approaches in several aspects. Although other methods have been proposed that estimate separately the biological and technical variance components (Daly, 2008; Clough, 2009), to the best of our knowledge, and with the exception of its predecessor (Jorge, 2009), no models have been formulated previously for quantitative proteomics data that decompose technical variance into two or more components. Besides, in our model the averages are calculated following error propagation theory, so that the statistical weight with which each value contributes to the average is exactly the inverse of its local variance, and the variances of each one of the averaged values are known with accuracy. A similar kind of weighting by the error made at the spectrum and peptide levels was proposed in one of the earliest models proposed to analyse quantitative data produced by using stable isotope-coded affinity tags (ICAT) (Li, 2003); however, the approach followed in that work only propagates the error produced at the time of quantification from the MS spectrum, and does not take into account the variance introduced by further errors produced by peptide generation and protein manipulation. Finally, to the best of our knowledge, our weighting method is the first one that has been demonstrated to be of general validity for a wide range of different SIL and MS approaches.

2.5 The standardised variable and the meaning of the outliers

The WSPP model also resolves the problem of undersampling and at the same time provides a framework to integrate data and a robust algorithm to estimate variances, which are of general applicability. This is accomplished by using standardised variables at each one of the integration levels. These variables express $\log_2$-ratios in units of standard deviation, introducing a bias correction for the number of degrees of freedom, which depends on the number of elements that are used to compute the average (such as the number of peptides that are used to estimate a protein value). Using standardised variables, all the elements at a given level of integration can be analysed together in a unique distribution that, under the null hypothesis, is expected to be a normal distribution with unit variance. We take advantage of this general property to make robust estimates of variances by an iterative method that is of general applicability for all integration levels and quantification approaches. We should note here that our approach to estimate variance is similar to that used by other authors (Cox and Mann, 2008; Zhang, 2010; Arntzen, 2011), although in these works only the total technical variance was estimated. The analysis of the standardised variable is also very useful to detect the presence of outliers at any integration level; in the WSPP model, this may be done at seven different levels, and in each one it provides specific information. Thus, at the spectrum and peptide levels, outliers are indicative of incorrect quantifications produced by a variety of causes (for instance peak coelution, bad fitting or methionine oxidation (Jorge, 2009; Bonzon-Kulichenko, 2011b)), as commented above. At the protein level outliers indicate statistically-significant abundance changes. But, at other levels of integration, an outlier may also indicate that an experiment replicate gives quantitative results that deviate from the other replicates more than expected by chance alone, and the absence of outliers that all the replicates behave as expected by the null hypothesis (as observed in this work). We should note here that this global conception of variances, which considers together the whole wealth of data, is in contrast with other approaches commonly followed to detect outliers and artefacts—like Dixon's Q test (Li, 2003)—that analyse locally the distribution of the elements used to calculate a particular average, and that have a very limited utility when the number of elements is very low (i.e., when a protein is quantified by only two peptides). In the WSPP model all the elements, even single-hits, are assigned a local variance and this is done on the basis of only four parametres, which are estimated from the analysis of the whole collection of data.

3. The Systems Biology Triangle (SBT): a new philosophy to interpret proteome-based systems biology

3.1 A model for systems biology based on the coordinated behaviour of proteins

In this third and last published paper we present the SBT, which takes advantage of the Generic Integration Algorithm (GIA), a generalisation of the WSPP statistical model presented in the second work (see section 4), to develop an innovative approach to systems biology based on the unique properties of the dynamic of proteomes.

The SBT is the first algorithm to analyse the coordinated protein behaviour in high-throughput quantitative pairwise proteomics experiments. Taking into account this coordinated protein behaviour, the model is able to detect changes in groups of proteins that have been classified into functional categories. This quality contrasts with other approaches used to study protein coordination, such as those based on correlation analysis; unlike them, our approach takes information from ontological databases to classify proteins into functional categories. The SBT directly addresses the question whether a protein has a coordinated behaviour within a category by determining if it has the same relative abundance change as the other proteins in the category. Significant changes at the category level are then detected using only the population of proteins that have a coordinated behaviour in each category. Consistently, we define the degree of coordination as the fraction of changing proteins that are coordinated, including in the calculation the proteins that belong to changing categories (as shown in Figure 1E of third publication, page 66). This approach allows the analysis of protein coordination when only two conditions are compared, making it possible to measure the evolution and consistency of coordination over time. Here we demonstrate that this approach can capture coordinated protein behaviour in seven different biological models subjected to various kinds of perturbations. However, in our extensive experience with the model we have successfully detected functional changes due to coordinated action of proteins in hundreds of comparative pairwise experiments. For this reason, we believe that the SBT model provides the proteomics community with a very useful tool for biological interpretation of pairwise quantitative proteomics experiments, sparing us the need to calculate correlations in large sample datasets.

Since our model needs previous ontological knowledge about the quantified proteins, it is expected to be very sensitive to the quality and exhaustiveness of the classification. Indeed, we found a higher degree of coordination in databases using a more exhaustive classification method, such as PANTHER (Thomas, 2003; Mi, 2005) and KEGG (Kanehisa and Goto, 2000), while larger, less curated databases, tended to detect a higher number of coordinated proteins and categories, but with a lower degree of coordination (Figure 2). We also found that inclusion of information about protein complexes, such as the CORUM database (Ruepp, 2008), was particularly useful for detecting coordination, in agreement with the evidence that subunits of complexes have a remarkable tendency to be co-regulated (Carmi, 2006; Carmi, 2009; Hansson,

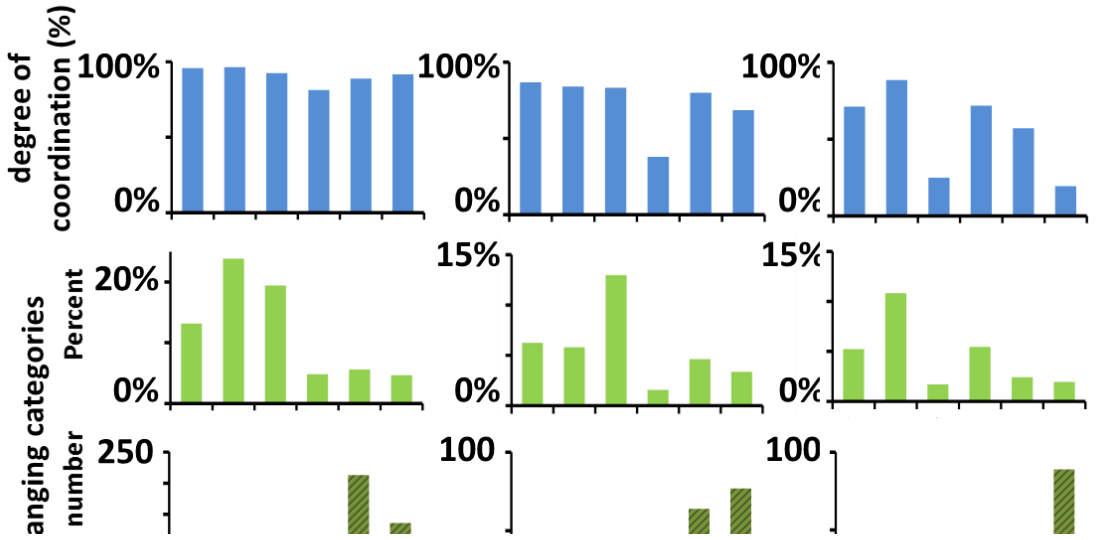


Figure 2: (Supplemental Figure S7 in the third paper (Garcia-Marques, 2016)**): Effect of the functional database on the degree of coordination and the sensitivity to detect coordinated category alterations**. Quantitative data from three different experiments were analysed using the SBT model using the functional databases shown. As it can be seen, for IPA and GO databases we can find more categories changing, although with a lower degree of coordination. On the other hand, PANTHER, KEGG and Reactome display less changing categories using absolute figures, but more in relation to the total number of categories managed, and with remarkably higher coordination.

2012). Interestingly, as it is described in the PhD Thesis from Fernando García-Marqués and published in the third article presented here (Garcia-Marques, 2016) we found that most of the outliers analysed here seem to play a differential role, like regulatory or signalling functions, suggesting that our algorithm can be used to detect protein abundance changes of particular biological relevance. Our finding resembles results obtained in a previous report, showing that deviation from the correlation pattern by one or more proteins indicated a specific function during the biological process (Hansson, 2012).

3.2 The contributions of the Systems Biology Triangle

The SBT model takes into account the dynamic range of protein abundances by introducing statistical weights when performing data integration, and also addresses the problem of undersampling. The model allows the detection of *protein to category* outliers, so that analysis of categories after outlier removal allows the detection of significant category changes that are produced by proteins acting in a coordinated manner, and hence can also be used as a robust functional class scoring algorithm. These properties of SBT make this model unique to treat quantitative proteomics data and to detect functional alterations that take place in a coordinated manner. These properties are not shared by other algorithms like GOrilla (Eden, 2009), STRING (Szklarczyk, 2014) or GSEA (Subramanian, 2005), and this fact explains why in this work they were not as sensitive as the SBT model to detect the functional categories affected by AngII in VSMC.

We should also note here that the SBT model is based on a null hypothesis that has been experimentally demonstrated. In previous studies we showed that the *z* variable at different levels consistently follows the expected $N(0,1)$ distribution (Jorge, 2009; Bonzon-Kulichenko, 2011b; Navarro, 2014). Here, using two lines of evidence, we further demonstrate that, in the null hypothesis, the standardised variable at the category level (z_{ca}) also follows the standard normal distribution. First, we demonstrate the absence of category changes in a null hypothesis experiment comparing two preparations of untreated yeast cells. Second, we demonstrate the absence of category changes when *protein to category* assignations were randomised, a process that disrupts any underlying coordination at the category levels. Therefore, the category changes may be detected by robust estimates of statistical significance. Notably, our integration model assigns more weight to quantifications of higher quality, making it unnecessary to eliminate poor quantifications (Jorge, 2009; Navarro, 2014), as stated in subsection 2.4.

We demonstrate the discriminatory power of SBT by unravelling with unprecedented molecular detail the coordinated mechanisms taking place during AngII stimulation of VSMC (in the PhD thesis from Fernando García-Marqués, and in the third paper presented in this work). The robustness of the model is supported by remarkable maintenance of coordination in the same categories in different VSMC preparations over time. From our data we cannot judge whether this regulation occurs at transcriptional or posttranscriptional levels; however, we found that the vast majority of coordinated categories are protein complexes, metabolic pathways or proteins forming interaction networks, suggesting that protein-protein interactions play an important role in producing tight co-regulation of protein levels. This conclusion is supported by

detection of proteins with a seemingly uncoordinated behaviour at the earliest time points, while the coordinated response elicited by AngII in VSMC builds up gradually, suggesting an adjustment of the regulatory protein machinery.

4. An innovative conception for the automatic statistical analysis of quantitative proteomics experiments

4.1 The Generic Integration Algorithm (GIA)

The WSPP model (Navarro, 2014) inspired us to develop the Generic Integration Algorithm (GIA), which lies at the heart of the SBT model. The GIA represents a generalisation of the different levels in the WSPP statistical models. Instead of treating each level differently, with similar, although separated algorithms, the whole process was restructured in order to have only a single algorithm, and treating each integration as a particular case.

Hence, the GIA rigorously integrates errors made at the inferior level (which can be, for example, the dataset of spectra, or the peptides) with the error made at the superior level (whose dataset might be the peptides if the lower level are spectra, or the protein data, if the lower level are peptides), allowing full control of the error associated with the integrated elements (i.e., proteins or categories) so that quantitative information may be analysed using the standardised *z* variable (i.e., abundance changes expressed in units of standard deviations).

4.2 The software platform SanXoT

Compared to QuiXoT (Pedro Navarro, PhD Thesis), which was our previous software supporting de WSPP model, the GIA and its accompanying software SanXoT allowed the automatic removal of outliers, calculated the variance in a more reliable way, could be completely automated, and permitted an unprecedented flexibility for the structure of workflows. First, the automatic removal of outliers (which had to be removed manually in QuiXoT) was done by a separate algorithm, allowing a refined and automatic removal of outliers without distorting the original distribution. Second, the variance calculation is more reliable, because instead of using a sweep algorithm to find the best variance (which generated some difficulties to find the statistical variance) SanXoT incorporated the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963), which is a fitting algorithm starting with a seed variance, and then approaching in cycles to the best variance that guarantees the standardised variable

follows a standard normal deviation (even in the cases when the original seed variance is very far from the solution). Third, it was completely automated, as it was designed to be used in steps such that the output of a program could be the input of the next one, in a modular way, so that in most cases it did not require the attention of the user until the whole workflow was finished. Finally, the GIA and its related software (see Figure 3) supposed a paradigmatic change in the way proteomic samples were analysed in the laboratory, as it opened the former WSPP statistical model to limitless possibilities beyond the *scan>peptide>protein* Fundamental Workflow. Thus, the GIA can be used to construct compact workflows allowing the automatic analysis of redox proteomics experiments and experiment merging, among other approaches.

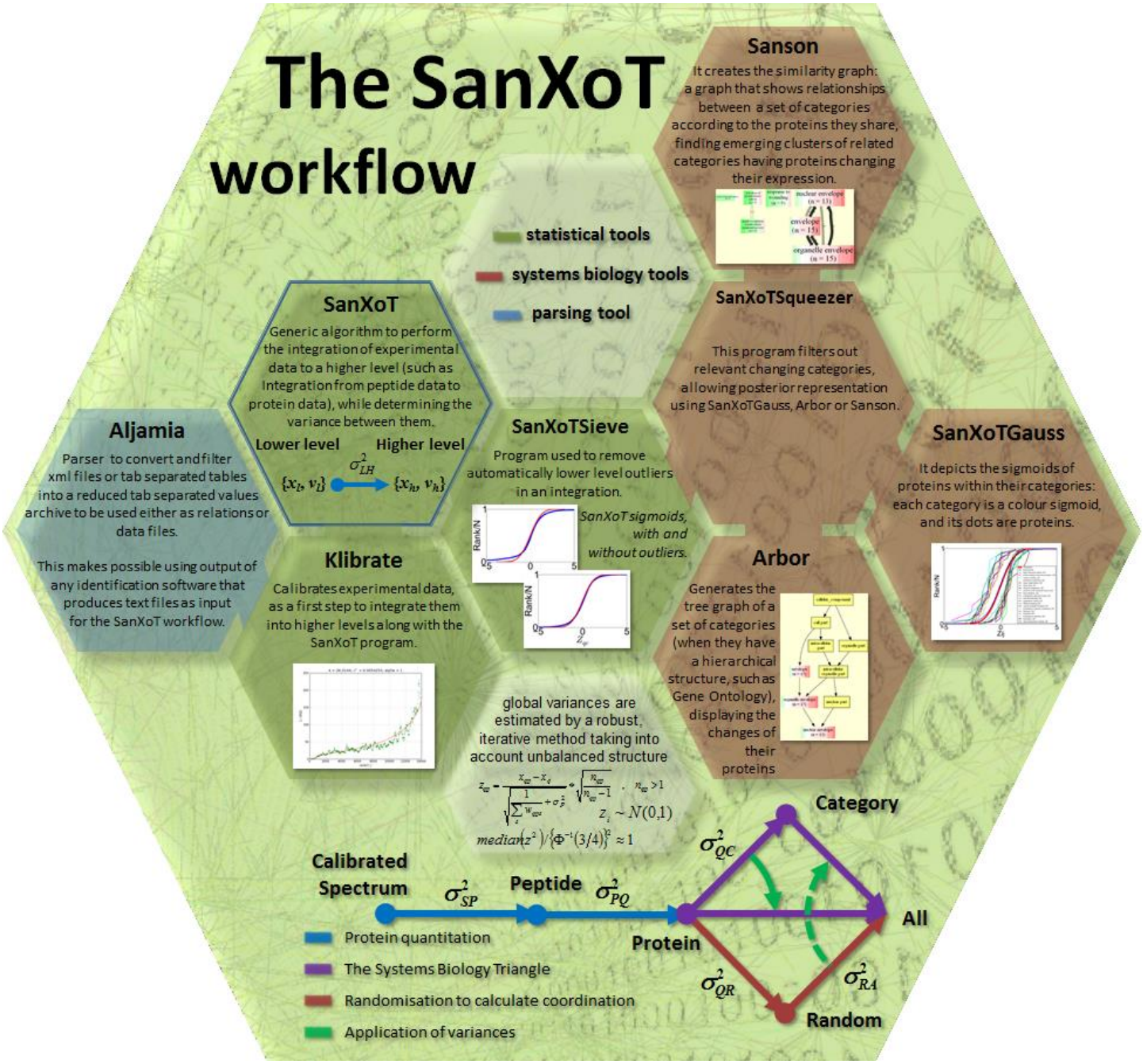


Figure 3: A summary of the different programs in the SanXoT workflow, and the interactions between them. At the bottom it is included the Fundamental Workflow assembled with the Systems Biology Triangle and the randomisation workflow to calculate the coordination. (Detail of awarded conference poster presented at the 13th Human Proteome World Congress, Madrid 2014, see full poster in Appendix 3).

5. Further research

5.1 Combination with the identification workflow under development

As already stated, the statistical model and its associated software have been designed keeping in mind the principle of maximum flexibility. One of the advantages is the possibility of combining quantitative proteomics workflows with an identification workflow currently under development. This identification workflow intends improving protein-level identification (instead of performing identification at the peptide level, as usual), accounting for the problem of distribution of peptides in different proteins. One possibility is merging both workflows in such a way that they both use the same peptide-to-protein relation tables, potentially allowing the improvement of the identification information by taking into account the information from quantitative proteomics experiments.

5.2 Analysis at peptide levels

The statistical model could also be potentially used to detect effects because of the presence of SNPs or differential splicing. In addition, the protein layer can also be ignored in our model so that statistical analysis of abundance changes is directly performed at the peptide level, without grouping peptides into proteins. Using this approach the model is currently being used in the ongoing research of computational models for the analysis of post-translational modifications (PTM) (Navratan Bagwan, PhD Thesis), which in turn, among other applications, is of great interest for the study of redox modifications, a large field with many implications that is under development in our laboratory.

5.3 Incorporation of multivariate analysis

The current model permits comparing two proteomic samples at once, which in some extent simplifies difficult problems by reducing the analysis always to a two-side comparison. For example, as it has been explained, it is possible combining different experiments into a single one: this is done by fusing data into a single reduced pairwise experiment (for example, *N* conditions vs *N* controls, can be integrated into a single "condition vs control", the same way the information of *N* peptides is combined into information of a single protein). In addition to this, it could be of interest exploring the potential of developing a version of the model incorporating multivariate analysis, especially for the analysis of multiplexed experiments such as iTRAQ or TMT.

5.4 Coordination in transcriptomics

In the discussion of the SBT (third paper), we said that we still do not know whether regulation leading to the protein coordinated behaviour originates at transcriptional or posttranscriptional levels. Hence, one interesting extension of this work could be using the coordination concept and the model, or a modification of it, to confirm or disprove that coordination arises already at transcriptomic level. During the elaboration of this work, a part of the effort has been made to detect coordination at transcriptional level using data from a microarray experiment, although the results (not shown) have been inconclusive. Nevertheless, some colleagues in the laboratory (Celia Castans, unpublished) have preliminary results applying the SBT model to data obtained by RNA-Seq (instead of microarrays), finding that the model explains very well the distribution of variances and suggesting that a high degree of coordination at the transcriptome level is very plausible.

5.5 Data independent acquisition (DIA) and label-free models

Another possible extension of this work could be using the model with data from DIA label-free experiments, in a similar way as it has been done with iTRAQ and TMT (taking as weight the intensity of the most intense element in a pairwise comparison). Nonetheless, this will require solving before statistical questions which might require an adjustment of the statistical model.

5.6 Newly developed SIL methods

The statistical model and the modular structure of the associated software—SanXoT—potentially allow a fast and easy adaptation to newly developed SIL methods. To illustrate an example about this, it might be worth mentioning NeuCode (Hebert, 2013): this is a SIL technique that takes advantage of the subtle mass differences due to ^{13}C and ^{15}N isotopes, combined with the increasing resolution of new instruments, to allow the advantages of multiplexing combined with an MS^1 spectrum. The current statistical model could, potentially, analyse a NeuCode experiment by adding a level (the "feature" level) prior to the scan level. This way, quantification of proteins would be possible by integrating different features into scan-level information, and then proceeding with the Fundamental Workflow as usual. Indeed, Salvador Martínez-Bartolomé, currently at the laboratory of John R. Yates, III (The Scripps Research Institute) is using the SanXoT workflow to treat data quantified employing a similar approach that uses reductive methylation for isobaric isotopologue labelling of peptides (Bamberger, 2014) with very promising results (*manuscript submitted*).

5.7 Parallelisation of software tools

The growing amount of data used in high-throughput proteomics increases, in turn, the amount of time and resources needed to complete the workflows. This adds pressure to further develop and improve the current structure of the algorithms. A study on the time complexity and granularity of the different parts of the algorithm can be performed to check elements with a high potential to be revised and parallelised. The current modular structure of the SanXoT software package facilitates these potential changes, although preserving this philosophy might be a challenge if deep modifications are needed.

5.8 Combination with network analysis

A considerable amount of research in proteomics consists in the development of network analysis techniques, such as those based on protein-protein interactions like STRING (Szklarczyk, 2014). A number of network analysis algorithms are currently available, such as the Molecular Complex Detection (MCODE) (Bader and Hogue, 2003), the Markov Clustering Algorithm (MCL) (Enright, 2002), the Restricted Neighborhood Search Clustering (RNSC) (King, 2004), or the Super paramagnetic clustering (SPC) (Blatt, 1996). An interesting possibility could be combining the SBT along with those network analysis algorithms to explore the potential of improving existing networks with data from quantitative experiments.

Conclusions

1. In peptide identification, using the decoy-target approach, and especially when scores are generated taking information from additional sequence candidates, the high mass accuracy of modern mass spectrometres should be exploited by combining wide precursor ion mass search windows followed by post-scoring mass filtering algorithms.

2. The WSPP statistical model (and its supporting software QuiXoT) provides a general statistical framework for high-throughput quantitative proteomics experiments, allowing the systematic comparison and integration of data from different SIL techniques. It allows the separation of the different sources of variance, enabling the interpretation of the random error at the different levels. Besides of its general applicability, the WSPP performance is similar or superior to other commonly employed methods.

3. The Systems Biology Triangle (SBT), based on the GIA, represents the first algorithm capable of analysing the coordinated behaviour of proteins. It permits the detection of functional categories affected by this behaviour, helping to interpret large-scale, high-throughput quantitative proteomics pairwise experiments. Furthermore, the SBT algorithm allows detecting the specific differential role of outlier proteins respect to other proteins classified in the same group, unveiling a wealth of biologically relevant regulatory or signalling information.

4. The Generic Integration Algorithm (GIA) and its accompanying software package (SanXoT) are an important technological advancement of the WSPP model. They can be used to construct compact workflows to integrate the information from quantitative proteomics experiments in a flexible fashion. Its applications include, but are not limited to, the use of the WSPP statistical model, the integration of SIL experiments of different nature, and the implementation of the Systems Biology Triangle.

Conclusiones

1. Al utilizar la estrategia de la base de datos señuelo (*target-decoy*) para identificar péptidos, y en concreto cuando las puntuaciones (*scores*) se generan empleando información de secuencias candidatas adicionales, la alta precisión de los espectrómetros de masas actuales se debería aprovechar combinando una ventana ancha para la masa los iones precursores seguida de un filtrado de dicha masa tras calcular la puntuación asignada a la identificación.

2. El modelo estadístico WSPP (del inglés *Weighted Spectrum, Peptide and Protein*), así como su *software* asociado, QuiXoT, proporcionan un marco estadístico general para proteómica cuantitativa de alto rendimiento, permitiendo la comparación sistemática y la integración de datos de diferentes técnicas de marcaje isotópico estable (SIL). Permite la separación de las diferentes fuentes de varianza, haciendo posible la interpretación del error aleatorio a diferentes niveles. Además de su aplicabilidad general, el desempeño del modelo WSPP es similar o superior al de otros modelos en uso.

3. El Triángulo de la Biología de Sistemas (SBT), basado en el Algoritmo de Integración Genérico (GIA), representa el primer algoritmo capaz de analizar el comportamiento coordinado de proteínas. Permite la detección de categorías funcionales influidas por estas pautas, ayudando a la interpretación a gran escala de experimentos binarios de proteómica cuantitativa de alto rendimiento. Asimismo, el algoritmo SBT proporciona los recursos necesarios para detectar las funciones diferenciales de proteínas de comportamiento atípico con respecto a otras clasificadas en el mismo grupo, revelando gran cantidad de información biológicamente relevante sobre funciones reguladoras o de señalización.

4. El Algoritmo de Integración Genérico (GIA), así como el paquete de *software* al que está asociado (SanXoT), suponen un importante avance tecnológico para el modelo WSPP. Pueden utilizarse para ensamblar flujos de trabajo compactos, permitiendo integrar la información de experimentos de proteómica cuantitativa con flexibilidad. Sus aplicaciones incluyen, entre otras, el uso del modelo WSPP, la integración de experimentos de marcaje isotópico estable (SIL) de distinta naturaleza, y la implementación del Triángulo de la Biología de Sistemas.

References

We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover all the tracks, to not worry about the blind alleys or to describe how you had the wrong idea first, and so on. So there isn't any place to publish, in a dignified manner, what you actually did in order to get to do the work

Richard Feynman, beginning of the Noble Prize lecture

Aderem, A. (2005) Systems biology: its practice and challenges. *Cell*, 121, 511-513.

Aebersold, R. (2009) A stress test for mass spectrometry-based proteomics. *Nat Methods*, 6, 411-412.

Ananiadou, S., Kell, D.B. and Tsujii, J.-i. (2006) Text mining and its potential applications in systems biology. *Trends in biotechnology*, 24, 571-579.

Arntzen, M.O., Koehler, C.J., Barsnes, H., Berven, F.S., Treumann, A. and Thiede, B. (2011) IsobariQ: software for isobaric quantitative proteomics using IPTL, iTRAQ, and TMT. *J Proteome Res*, 10, 913-920.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S. and Eppig, J.T. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, 25, 25-29.

Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4, 1.

Bamberger, C., Pankow, S., Park, S.K.R. and Yates III, J.R. (2014) Interference-free proteome quantification with MS/MS-based isobaric isotopologue detection. *Journal of proteome research*, 13, 1494-1501.

Bantscheff, M., Boesche, M., Eberhard, D., Matthieson, T., Sweetman, G. and Kuster, B. (2008) Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol Cell Proteomics*, 7, 1702-1713.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y. and Barkai, N. (2006) Noise in protein expression scales with natural protein abundance. *Nat Genet*, 38, 636-643.

Barabasi, A.-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5, 101-113.

Barry, W.T., Nobel, A.B. and Wright, F.A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21, 1943-1949.

Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. and Gygi, S.P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*, 24, 1285-1292.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.*, 1, 289-300.

Blatt, M., Wiseman, S. and Domany, E. (1996) Superparamagnetic clustering of data. *Physical review letters*, 76, 3251.

Boehm, A.M., Putz, S., Altenhofer, D., Sickmann, A. and Falk, M. (2007) Precise protein quantification based on peptide quantification using iTRAQ. *BMC bioinformatics*, 8, 214.

Bonzon-Kulichenko, E., Martinez-Martinez, S., Trevisan-Herraz, M., Navarro, P., Redondo, J.M. and Vazquez, J. (2011a) Quantitative in-depth analysis of the dynamic secretome of activated Jurkat T-cells. *Journal of proteomics*, 75, 561-571.

Bonzon-Kulichenko, E., Perez-Hernandez, D., Nunez, E., Martinez-Acedo, P., Navarro, P., Trevisan-Herraz, M., Ramos Mdel, C., Sierra, S., Martinez-Martinez, S., Ruiz-Meana, M., Miro-Casas, E., Garcia-Dorado, D., Redondo, J.M., Burgos, J.S. and Vazquez, J. (2011b) A robust method for quantitative high-throughput analysis of proteomes by 18O labeling. *Molecular & cellular proteomics : MCP*, 10, M110 003335.

Brosch, M., Swamy, S., Hubbard, T. and Choudhary, J. (2008) Comparison of Mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. *Mol Cell Proteomics*, 7, 962-970.

Burillo, E., Jorge, I., Martínez-López, D., Camafeita, E., Blanco-Colio, L.M., Trevisan-Herraz, M., Ezkurdia, I., Egido, J., Michel, J.-B., Meilhac, O., Vázquez, J. and Martin-Ventura, J.L. (2016) Quantitative HDL Proteomics Identifies Peroxiredoxin-6 as a Biomarker of Human Abdominal Aortic Aneurysm. *Scientific Reports*, 6, 38477.

Carmi, S., Levanon, E.Y., Havlin, S. and Eisenberg, E. (2006) Connectivity and expression in protein networks: Proteins in a complex are uniformly expressed. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 73, 031909-031906.

Carmi, S., Levanon, E.Y. and Eisenberg, E. (2009) Efficiency of complex production in changing environment. *BMC systems biology*, 3, 3.

Clough, T., Key, M., Ott, I., Ragg, S., Schadow, G. and Vitek, O. (2009) Protein quantification in label-free LC-MS experiments. *J Proteome Res*, 8, 5275-5284.

Cooper, B. (2011) The problem with peptide presumption and low Mascot scoring. *J Proteome Res*, 10, 1432-1435.

Cooper, B. (2012) The problem with peptide presumption and the downfall of target-decoy false discovery rates. *Anal Chem*, 84, 9663-9667.

Cottrell, J.S. and London, U. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *electrophoresis*, 20, 3551-3567.

Cottrell, J.S. and Creasy, D.M. (2011) Response to: the problem with peptide presumption and low Mascot scoring. *J Proteome Res*, 10, 5272-5273.

Cox, J. and Mann, M. (2007) Is proteomics the new genomics? *Cell*, 130, 395-398.

Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26, 1367-1372.

Cox, J. and Mann, M. (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry*, 80, 273-299.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V. and Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. In, *J. Proteome Res.*; 2011. p. 1794-1805.

Craig, R. and Beavis, R.C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid communications in mass spectrometry*, 17, 2310-2316.

Craig, R. and Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20, 1466-1467.

Chalkley, R.J. (2013) When target-decoy false discovery rate estimations are inaccurate and how to spot instances. *J Proteome Res*, 12, 1062-1064.

Chang, C.Y., Picotti, P., Huttenhain, R., Heinzelmann-Schwarz, V., Jovanovic, M., Aebersold, R. and Vitek, O. (2012) Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol Cell Proteomics*, 11, M111 014662.

Daly, D.S., Anderson, K.K., Panisko, E.A., Purvine, S.O., Fang, R., Monroe, M.E. and Baker, S.E. (2008) Mixed-effects statistical model for comparative LC-MS proteomics studies. *J Proteome Res*, 7, 1209-1217.

Danielsen, M., Hornshoj, H., Siggers, R.H., Jensen, B.B., van Kessel, A.G. and Bendixen, E. (2007) Effects of bacterial colonization on the porcine intestinal proteome. *Journal of proteome research*, 6, 2596-2604.

Ding, Y., Choi, H. and Nesvizhskii, A.I. (2008) Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. *J Proteome Res*, 7, 4878-4889.

Domon, B. and Aebersold, R. (2010) Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol*, 28, 710-721.

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10, 1.

Elias, J.E. and Gygi, S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4, 207-214.

Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30, 1575-1584.

Ezkurdia, I., Calvo, E., Del Pozo, A., Vázquez, J., Valencia, A. and Tress, M.L. (2015) The potential clinical impact of the release of two drafts of the human proteome. *Expert review of proteomics*, 12, 579-593.

Finkelstein, A., Hetherington, J., Margoninski, O., Saffrey, P., Seymour, R. and Warner, A. (2004) Computational challenges of systems biology. *Computer*, 37, 26-33.

Fisher, R.A. The design of experiments. 1935.

Foster, L.J., de Hoog, C.L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V.K. and Mann, M. (2006) A mammalian organelle map by protein correlation profiling. *Cell*, 125, 187-199.

Gan, C.S., Chong, P.K., Pham, T.K. and Wright, P.C. (2007) Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). *J Proteome Res*, 6, 821-827.

Gandhi, S.J., Zenklusen, D., Lionnet, T. and Singer, R.H. (2011) Transcription of functionally related constitutive genes is not coordinated. *Nat Struct Mol Biol*, 18, 27-34.

Garcia-Marques, F., Trevisan-Herraz, M., Martinez-Martinez, S., Camafeita, E., Jorge, I., Lopez, J.A., Mendez-Barbero, N., Mendez-Ferrer, S., Del Pozo, M.A., Ibanez, B., Andres, V., Sanchez-Madrid, F., Redondo, J.M., Bonzon-Kulichenko, E. and Vazquez, J. (2016) A novel systems-biology algorithm for the analysis of coordinated protein responses using quantitative proteomics. *Mol Cell Proteomics*.

Garcia-Prieto, J., Garcia-Ruiz, J.M., Sanz-Rosa, D., Pun, A., Garcia-Alvarez, A., Davidson, S.M., Fernandez-Friera, L., Nuno-Ayala, M., Fernandez-Jimenez, R., Bernal, J.A., Izquierdo-Garcia, J.L., Jimenez-Borreguero, J., Pizarro, G., Ruiz-Cabello, J., Macaya, C., Fuster, V., Yellon, D.M. and Ibanez, B. (2014) beta3 adrenergic receptor selective stimulation during ischemia/reperfusion improves cardiac function in translational models through inhibition of mPTP opening in cardiomyocytes. *Basic research in cardiology*, 109, 422.

Ge, H., Walhout, A.J. and Vidal, M. (2003) Integrating 'omic'information: a bridge between genomics and systems biology. *TRENDS in Genetics*, 19, 551-560.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast. *Nature*, 425, 737-741.

Gonzalez-Granado, J.M., Silvestre-Roig, C., Rocha-Perugini, V., Trigueros-Motos, L., Cibrian, D., Morlino, G., Blanco-Berrocal, M., Osorio, F.G., Freije, J.M., Lopez-Otin, C., Sanchez-Madrid, F. and Andres, V. (2014) Nuclear envelope lamin-A couples actin dynamics with immunological synapse architecture and T cell activation. *Science signaling*, 7, ra37.

Hansson, J., Rafiee, M.R., Reiland, S., Polo, J.M., Gehring, J., Okawa, S., Huber, W., Hochedlinger, K. and Krijgsveld, J. (2012) Highly Coordinated Proteome Dynamics during Reprogramming of Somatic Cells to Pluripotency. *Cell reports*, 2, 1579-1592.

Hebert, A.S., Merrill, A.E., Bailey, D.J., Still, A.J., Westphall, M.S., Strieter, E.R., Pagliarini, D.J. and Coon, J.J. (2013) Neutron-encoded mass signatures for multiplexed proteome quantification. *Nat Methods*, 10, 332-334.

Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C. and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A*, 90, 5011-5015.

Herbrich, S.M., Cole, R.N., West, J., Keith P, Schulze, K., Yager, J.D., Groopman, J.D., Christian, P., Wu, L., O'Meally, R.N., May, D.H., McIntosh, M.W. and Ruczinski, I. (2013) Statistical Inference from Multiple iTRAQ Experiments without Using Common Reference Standards. *Journal of Proteome Research*, 12, 594-604.

Hill, E.G., Schwacke, J.H., Comte-Walters, S., Slate, E.H., Oberg, A.L., Eckel-Passow, J.E., Therneau, T.M. and Schey, K.L. (2008) A statistical model for iTRAQ data analysis. *J Proteome Res*, 7, 3091-3101.

Hood, L., Heath, J.R., Phelps, M.E. and Lin, B. (2004) Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306, 640-643.

Hsieh, E.J., Hoopmann, M.R., MacLean, B. and MacCoss, M.J. (2010) Comparison of database search strategies for high precursor mass accuracy MS/MS data. *Journal of proteome research*, 9, 1138-1143.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37, 1-13.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1, S96-104.

Hume, D. A Treatise of Human Nature. 1738.

Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2, 343-372.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002) Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31, 370-377.

Ihmels, J., Levy, R. and Barkai, N. (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nat Biotechnol*, 22, 86-92.

Isern, J., Martin-Antonio, B., Ghazanfari, R., Martin, A.M., Lopez, J.A., del Toro, R., Sanchez-Aguilera, A., Arranz, L., Martin-Perez, D., Suarez-Lledo, M., Marin, P., Van Pel, M., Fibbe, W.E., Vazquez, J., Scheding, S., Urbano-Ispizua, A. and Mendez-Ferrer, S. (2013) Self-renewing human bone marrow mesenspheres promote hematopoietic stem cell expansion. *Cell reports*, 3, 1714-1724.

James, P., Quadroni, M., Carafoli, E. and Gonnet, G. (1993) Protein identification by mass profile fingerprinting. *Biochem Biophys Res Commun*, 195, 58-64.

Jeong, K., Kim, S. and Bandeira, N. (2012) False discovery rates in spectral identification. *BMC Bioinformatics*, 13, 1.

Jiang, Z. and Gentleman, R. (2007) Extensions to gene set enrichment. *Bioinformatics*, 23, 306-313.

Jones, A.R., Siepen, J.A., Hubbard, S.J. and Paton, N.W. (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics*, 9, 1220-1229.

Jorge, I., Navarro, P., Martinez-Acedo, P., Nunez, E., Serrano, H., Alfranca, A., Redondo, J.M. and Vazquez, J. (2009) Statistical model to analyze quantitative proteomics data obtained by 18O/16O labeling and linear ion trap mass spectrometry: application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells. *Molecular & cellular proteomics : MCP*, 8, 1130-1149.

Jorge, I., Burillo, E., Mesa, R., Baila-Rueda, L., Moreno, M., Trevisan-Herraz, M., Silla-Castro, J.C., Camafeita, E., Ortega-Muñoz, M., Bonzon-Kulichenko, E., Calvo, I., Cenarro, A., Civeira, F. and Vázquez, J. (2014) The human HDL proteome displays high inter-individual variability and is altered dynamically in response to angioplasty-induced atheroma plaque rupture. *Journal of proteomics*, 106, 61-73.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G. and Matthews, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33, D428-D432.

Käll, L., Storey, J.D., MacCoss, M.J. and Noble, W.S. (2007) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, 7, 29-34.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28, 27-30.

Kanehisa, M. (2016) KEGG Bioinformatics Resource for Plant Genomics and Metabolomics. *Plant Bioinformatics: Methods and Protocols*, 55-70.

Karp, N.A., Huber, W., Sadowski, P.G., Charles, P.D., Hester, S.V. and Lilley, K.S. (2010) Addressing accuracy and precision issues in iTRAQ quantitation. *Mol Cell Proteomics*, 9, 1885-1897.

Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J.N., Ansong, C., Heffron, F., Metz, T.O., Qian, W.J., Yoon, H., Smith, R.D. and Dabney, A.R. (2009) A

statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, 25, 2028-2034.

Kell, D.B. (2004) Metabolomics and systems biology: making sense of the soup. *Current opinion in microbiology*, 7, 296-307.

Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21, 3587-3595.

Khatri, P., Sirota, M. and Butte, A.J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. In, *PLoS Comput Biol*. 2012. p. e1002375.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., Thomas, J.K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N.A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L.D., Patil, A.H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S.K., Marimuthu, A., Sathe, G.J., Chavan, S., Datta, K.K., Subbannayya, Y., Sahu, A., Yelamanchi, S.D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K.R., Syed, N., Goel, R., Khan, A.A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.C., Zhong, J., Wu, X., Shaw, P.G., Freed, D., Zahari, M.S., Mukherjee, K.K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C.J., Shankar, S.K., Satishchandra, P., Schroeder, J.T., Sirdeshmukh, R., Maitra, A., Leach, S.D., Drake, C.G., Halushka, M.K., Prasad, T.S., Hruban, R.H., Kerr, C.L., Bader, G.D., Iacobuzio-Donahue, C.A., Gowda, H. and Pandey, A. (2014) A draft map of the human proteome. *Nature*, 509, 575-581.

King, A.D., Pržulj, N. and Jurisica, I. (2004) Protein complex prediction via cost-based clustering. *Bioinformatics*, 20, 3013-3020.

Kirschner, M.W. (2005) The meaning of systems biology. *Cell*, 121, 503-504.

Kitano, H. Foundations of systems biology. MIT press Cambridge; 2001.

Kitano, H. (2002) Systems biology: a brief overview. *Science*, 295, 1662-1664.

Latorre-Pellicer, A., Moreno-Loshuertos, R., Lechuga-Vieco, A.V., Sánchez-Cabo, F., Torroja, C., Acín-Pérez, R., Calvo, E., Aix, E., González-Guerra, A. and Logan, A. (2016) Mitochondrial and nuclear DNA matching shapes metabolism and healthy ageing. *Nature*, 535, 561-565.

Levenberg, K. (1944) A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2, 164-168.

Li, X.J., Zhang, H., Ranish, J.A. and Aebersold, R. (2003) Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Analytical chemistry*, 75, 6648-6657.

Lin, W.T., Hung, W.N., Yian, Y.H., Wu, K.P., Han, C.L., Chen, Y.R., Chen, Y.J., Sung, T.Y. and Hsu, W.L. (2006) Multi-Q: a fully automated tool for multiplexed protein quantitation. *J Proteome Res*, 5, 2328-2338.

Mahoney, D.W., Therneau, T.M., Heppelmann, C.J., Higgins, L., Benson, L.M., Zenka, R.M., Jagtap, P., Nelsestuen, G.L., Bergen, I., H Robert and Oberg, A.L. (2011) Relative Quantification: Characterization of Bias, Variability and Fold Changes in Mass Spectrometry Data from iTRAQ-Labeled Peptides. *Journal of Proteome Research*, 10, 4325-4333.

Mann, M., Hojrup, P. and Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom*, 22, 338-345.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R. and Bähler, J. (2012) Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell*, 151, 671-683.

Marquardt, D.W. (1963) An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11, 431-441.

Martinez-Acedo, P., Nunez, E., Gomez, F.J., Moreno, M., Ramos, E., Izquierdo-Alvarez, A., Miro-Casas, E., Mesa, R., Rodriguez, P., Martinez-Ruiz, A., Dorado, D.G., Lamas, S. and Vazquez, J. (2012) A novel strategy for global analysis of the dynamic thiol redox proteome. *Mol Cell Proteomics*, 11, 800-813.

Martinez-Bartolome, S., Navarro, P., Martin-Maroto, F., Lopez-Ferrer, D., Ramos-Fernandez, A., Villar, M., Garcia-Ruiz, J.P. and Vazquez, J. (2008) Properties of average score distributions of SEQUEST: the probability ratio method. *Mol Cell Proteomics*, 7, 1135-1145.

Martínez-Bartolomé, S., Navarro, P., Martín-Maroto, F., López-Ferrer, D., Ramos-Fernández, A., Villar, M., García-Ruiz, J.P. and Vázquez, J. (2008) Properties of Average Score Distributions of SEQUEST The Probability Ratio Method. *Molecular & Cellular Proteomics*, 7, 1135-1145.

Marx, V. (2013) Biology: The big challenges of big data. *Nature*, 498, 255-260.

Mason, O. and Verwoerd, M. (2007) Graph theory and networks in biology. *IET systems biology*, 1, 89-119.

Mateos-Hernández, L., Villar, M., Doncel-Pérez, E., Trevisan-Herraz, M., García-Forcada, Á., Romero, G.F., Vázquez, J. and Fuente, J. (2016) Quantitative proteomics reveals Piccolo as a candidate serological correlate of recovery from Guillain-Barré syndrome. *Oncotarget*.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., Kitano, H. and Thomas, P.D. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*, 33, D284-D288.

Mirgorodskaya, O.A., Kozmin, Y.P., Titov, M.I., Korner, R., Sonksen, C.P. and Roepstorff, P. (2000) Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. *Rapid Commun Mass Spectrom*, 14, 1226-1232.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34, 267-273.

Mortz, E., O'Connor, P.B., Roepstorff, P., Kelleher, N.L., Wood, T.D., McLafferty, F.W. and Mann, M. (1996) Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases. *Proc Natl Acad Sci U S A*, 93, 8264-8267.

Navarro, P. and Vazquez, J. (2009) A refined method to calculate false discovery rates for peptide identification using decoy databases. *Journal of proteome research*, 8, 1792-1796.

Navarro, P., Trevisan-Herraz, M., Bonzon-Kulichenko, E., Nunez, E., Martinez-Acedo, P., Perez-Hernandez, D., Jorge, I., Mesa, R., Calvo, E., Carrascal, M., Hernaez, M.L., Garcia, F., Barcena, J.A., Ashman, K., Abian, J., Gil, C., Redondo, J.M. and Vazquez, J. (2014) General statistical framework for quantitative proteomics by stable isotope labeling. *Journal of proteome research*, 13, 1234-1247.

Nesvizhskii, A.I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*, 73, 2092-2123.

Newman, J., Ghaemmaghami, S. and Ihmels, J. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. In, *Nature*. 2006.

Nilsson, T., Mann, M., Aebersold, R., Yates, J.R., 3rd, Bairoch, A. and Bergeron, J.J. (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods*, 7, 681-685.

Oberg, A.L., Mahoney, D.W., Eckel-Passow, J.E., Malone, C.J., Wolfinger, R.D., Hill, E.G., Cooper, L.T., Onuma, O.K., Spiro, C., Therneau, T.M. and Bergen, I., H Robert (2008) Statistical Analysis of Relative Labeled Mass Spectrometry Data from Complex Samples Using ANOVA. *Journal of Proteome Research*, 7, 225-233.

Oberg, A.L. and Vitek, O. (2009) Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res*, 8, 2144-2156.

Oberg, A.L. and Mahoney, D.W. (2012) Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. *BMC Bioinformatics*, 13 Suppl 16, S7.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A. and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1, 376-386.

Ong, S.E. and Mann, M. (2006) A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc*, 1, 2650-2660.

Pappin, D.J., Hojrup, P. and Bleasby, A.J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol*, 3, 327-332.

Pearson, K. (1900) X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50, 157-175.

Polpitiya, A.D., Qian, W.J., Jaitly, N., Petyuk, V.A., Adkins, J.N., Camp, D.G., 2nd, Anderson, G.A. and Smith, R.D. (2008) DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics*, 24, 1556-1558.

Popper, K. (1963) Conjectures and refutations. *The Growth of Scientific Knowledge*.

Rodriguez-Suarez, E., Gubb, E., Alzueta, I.F., Falcon-Perez, J.M., Amorim, A., Elortza, F. and Matthiesen, R. (2010) Virtual expert mass spectrometrist: iTRAQ tool for database-dependent search, quantitation and result storage. *Proteomics*, 10, 1545-1556.

Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom*, 11, 601.

Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A. and Pappin, D.J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*, 3, 1154-1169.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V. and Mewes, H.W. (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, 36, D646-D650.

Sagan, C., *Cosmos: Encyclopaedia Galactica (episode 12)*, 1980 (video), min: 1:10.

Shadforth, I.P., Dunkley, T.P., Lilley, K.S. and Bessant, C. (2005) i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics*, 6, 145.

Sprinzak, E., Cokus, S.J., Yeates, T.O., Eisenberg, D. and Pellegrini, M. (2009) Detecting coordinated regulation of multi-protein complexes using logic analysis of gene expression. *BMC Syst Biol*, 3, 115.

Stelling, J. (2004) Mathematical models in microbial systems biology. *Current opinion in microbiology*, 7, 513-518.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102, 15545-15550.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A. and Tsafou, K.P. (2014) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, gku1003.

Tanay, A., Sharan, R., Kupiec, M. and Shamir, R. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS*, 101, 2981-2986.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*, 13, 2129-2141.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T. and Hamon, C. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry*, 75, 1895-1904.

Tomita, M. and Kami, K. (2012) Systems biology, metabolomics, and cancer metabolism. *Science*, 336, 990-991.

Unwin, R.D., Pierce, A., Watson, R.B., Sternberg, D.W. and Whetton, A.D. (2005) Quantitative proteomic analysis using isobaric protein tags enables rapid comparison of changes in transcript and protein levels in transformed cells. *Mol Cell Proteomics*, 4, 924-935.

Varela, I., Cadinanos, J., Pendas, A.M., Gutierrez-Fernandez, A., Folgueras, A.R., Sanchez, L.M., Zhou, Z., Rodriguez, F.J., Stewart, C.L., Vega, J.A., Tryggvason, K., Freije, J.M. and Lopez-Otin, C. (2005) Accelerated ageing in mice deficient in Zmpste24 protease is linked to p53 signalling activation. *Nature*, 437, 564-568.

Weckwerth, W. (2003) Metabolomics in systems biology. *Annual review of plant biology*, 54, 669-689.

Wei, H., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G.P., Somerville, C. and Loraine, A. (2006) Transcriptional Coordination of the Metabolic Network in Arabidopsis. *Plant Physiology*, 142, 762-774.

Westerhoff, H.V. and Palsson, B.O. (2004) The evolution of molecular biology into systems biology. *Nature biotechnology*, 22, 1249-1252.

Weston, A.D. and Hood, L. (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of proteome research*, 3, 179-196.

Wieckowski, M.R., Giorgi, C., Lebiedzinska, M., Duszynski, J. and Pinton, P. (2009) Isolation of mitochondria-associated membranes and mitochondria from animal tissues and cells. *Nature protocols*, 4, 1582-1590.

Wiener, N. Cybernetics: Control and communication in the animal and the machine. Wiley New York; 1948.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-

Huspenina, J., Boese, J.H., Bantscheff, M., Gerstmair, A., Faerber, F. and Kuster, B. (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, 509, 582-587.

Wu, Y., Williams, E.G., Dubuis, S., Mottis, A., Jovaisaite, V., Houten, S.M., Argmann, C.A., Faridi, P., Wolski, W., Kutalik, Z., Zamboni, N., Auwerx, J. and Aebersold, R. (2014) Multilayered Genetic and Omics Dissection of Mitochondrial Activity in a Mouse Reference Population. *Cell*, 158, 1415-1430.

Yates, J.R., 3rd, Speicher, S., Griffin, P.R. and Hunkapiller, T. (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem*, 214, 397-408.

Yates, J.R., 3rd, Eng, J.K. and McCormack, A.L. (1995a) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*, 67, 3202-3210.

Yates, J.R., 3rd, Eng, J.K., McCormack, A.L. and Schieltz, D. (1995b) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, 67, 1426-1436.

Zhang, Y., Askenazi, M., Jiang, J., Luckey, C.J., Griffin, J.D. and Marto, J.A. (2010) A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. *Mol Cell Proteomics*, 9, 780-790.

Appendices

Appendix 1: Other papers to which this work has contributed directly

1.1 The human HDL proteome displays high inter-individual variability and is altered dynamically in response to angioplasty-induced atheroma plaque rupture

(Information on this publication is in subsection 4.2 of the Results chapter.)

JOURNAL OF PROTEOMICS 106 (2014) 61–73

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

The human HDL proteome displays high inter-individual variability and is altered dynamically in response to angioplasty-induced atheroma plaque rupture

Inmaculada Jorge[a,b,1], *Elena Burillo*[a,c,1], *Raquel Mesa*[a,b], *Lucía Baila-Rueda*[c], *Margoth Moreno*[a], *Marco Trevisan-Herraz*[a,b], *Juan Carlos Silla-Castro*[a,b], *Emilio Camafeita*[a,b], *Mariano Ortega-Muñoz*[a], *Elena Bonzon-Kulichenko*[a,b], *Isabel Calvo*[d], *Ana Cenarro*[c], *Fernando Civeira*[c], *Jesús Vázquez*[a,b,*]

[a]*Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Madrid, Spain*
[b]*Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain*
[c]*Hospital Universitario Miguel Servet, Instituto de Investigación Sanitaria de Aragón, I+CS, Zaragoza, Spain*
[d]*Hospital Universitario Miguel Servet, Servicio de Cardiología, Zaragoza, Spain*

ARTICLE INFO

Article history:
Received 23 December 2013
Accepted 7 April 2014
Available online 18 April 2014

Keywords:
HDL
Atheroma plaque
Proteomics
Stable isotopic labeling
System biology

ABSTRACT

Recent findings support potential roles for HDL in cardiovascular pathophysiology not related to lipid metabolism. We address whether HDL proteome is dynamically altered in atheroma plaque rupture. We used immunoaffinity purification of HDL samples from coronary artery disease patients before and after percutaneous transluminal coronary angioplasty (PTCA), a model of atheroma plaque disruption. Samples were analyzed by quantitative proteomics using stable isotope labeling and results were subjected to statistical analysis of protein variance using a novel algorithm. We observed high protein variability in HDL composition between individuals, indicating that HDL protein composition is highly patient-specific. However, intra-individual protein variances remained at low levels, confirming the reproducibility of the method used for HDL isolation and protein quantification. A systems biology analysis of HDL protein alterations induced by PTCA revealed an increase in two protein clusters that included several apolipoproteins, fibrinogen-like protein 1 and other intracellular proteins, and a decrease in antithrombin-III, annexin A1 and several immunoglobulins. Our results support the concept of HDL as dynamic platforms that donate and receive a variety of molecules and provide an improved methodology to use HDL proteome for the systematic analysis of differences among individuals and the search for cardiovascular biomarkers.

* *Corresponding author at:* Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Melchor Fernández Almagro, 3. 28029, Madrid, Spain.
E-mail address: jvazquez@cnic.es (J. Vázquez).
[1] These authors contributed equally in this work.

http://dx.doi.org/10.1016/j.jprot.2014.04.010

Biological significance

The HDL proteome is an interesting model of clinical relevance and has been previously described to be dynamically altered in response to pathophysiological conditions and cardiovascular diseases. Our study suggests that interindividual variability of HDL proteome is higher than previously thought and provided the detection of a set of proteins that changed their abundance in response to plaque rupture, supporting the concept of HDL as dynamic platforms that donate and receive a variety of molecules.

1. Introduction

Clinical, epidemiologic and genetic studies have confirmed the inverse relationship between high-density lipoprotein (HDL) cholesterol and coronary artery disease (CAD) [1,2]. However, although clinical trials have tested various approaches to increasing HDL concentration, the results of these trials have been mostly unsuccessful [3]. Moreover, two recent meta-analyses have concluded that increasing HDL cholesterol concentration does not reduce the risk of coronary disease events [4,5], suggesting that the well-established cardio-protective effects of HDL might be at least partly independent of its lipid content. In the specific case of CAD patients undergoing percutaneous transluminal coronary angioplasty (PTCA) and after reaching the therapeutic target of low LDL-cholesterol, an increase of HDL cholesterol demonstrated beneficial effects in cardiovascular risk [6].

Recent studies demonstrate that HDL particles have a very dynamic protein composition that is linked to innate immunity, endothelial vascular function, regulation of protease activity, oxidation, thrombosis, and inflammation [7–9]. These findings support potential roles for HDL particles in cardiovascular and atherosclerosis protection other than reverse cholesterol transport. It is noteworthy that the HDL proteome includes several serine proteinase inhibitors, called serpins [9]. Serpins are implicated in extracellular matrix degradation [10], and may be involved in the hypothesized role of HDL in supporting macrophage-mediated removal of apoptotic cells at sites of tissue damage and in the proposal that HDL and the vascular wall exchange not only excess cholesterol but also proteins related to tissue damage and infection [11]. In this view, HDL particles would serve as a platform for protein complexes that have specific functions related to atherogenesis and plaque rupture [12]. Based on this, we hypothesize that HDL participates in plaque rupture by modulating extracellular matrix degradation and facilitating reverse protein transport from the atherosclerotic plaque. Here, we address the question whether the protein composition of HDL in stable CAD patients is altered when the particle interacts with a ruptured atheroma plaque. We selected PTCA as a model of plaque rupture and used affinity chromatography to isolate HDL particles. The samples were analyzed by quantitative proteomics using stable isotope labeling and the results were subjected to statistical analysis of protein variance using a novel algorithm. We were able to study the variability in protein composition of HDL particles in different individuals and the pattern of changes in HDL protein composition as a result of PTCA. Our results are coherent with the dynamic nature of HDL particles and provide a set of methodologies for the systematic study of this proteome.

2. Material and methods

2.1. Subject selection and plasma collection

All participants gave written informed consent to a research protocol previously approved by the ethics committee of the Hospital Universitario Miguel Servet (Comité Ético de Investigación Clínica de Aragón, CEICA). The study was conducted in 21 male subjects with stable ischemic coronary disease (types I and II in the Canadian Cardiovascular Society classification). Patients were hemodynamically stable and had either >70% stenosis in at least two coronary vessels determined by coronary angiography, or, if stenosis was determined by intravascular ultrasound (IVUS), a flow-free area <6 mm^2 in the left common artery or <4 mm^2 in the other coronary arteries. In every case, at least two atherosclerotic plaques were treated by angioplasty during the procedure. The main clinical characteristics of these patients are summarized in Supplementary Excel Table 1. Coronary sinus blood was extracted from patients fasted for ≥12 h before and after the PTCA procedure using Vacutainer tubes with EDTA as anticoagulant. 50 mg of Heparin-Na was administrated to patients before procedure. Pre-PTCA samples were obtained from the catheter immediately before the catheterization of the left main coronary artery. Post-PTCA samples were obtained when the angioplasty procedure was finished, approximately 30–60 min after the first angioplasty once the catheter reached the coronary sinus during its retirement from the coronary tree. No systemic or coronary drugs, other than the dye, were used during the procedure. Plasma was separated from cellular components by centrifugation at 3000 rpm for 15 min at 4 °C, and 1.5 mM sodium azide and the 0.5 mM protease inhibitor Pefabloc SC (Roche) were added. All samples were stored at −80 °C until biochemical and proteomic analyses.

2.2. HDL isolation by immunoaffinity chromatography

Two milliliters of plasma from a single patient were applied to an anti-ApoA-I affinity column made with 5 ml HiTrap™ NHS-activated HP (GE Healthcare) coupled to 5 mg anti-human ApoA-I antibody (BioDesign International) and purified on an ÄKTA Purifier UPC 10 FPLC system (GE Healthcare) at a flow rate of 1 ml/min with a maximum pressure of 0.5 MPa [13]. The column was previously equilibrated with 10 ml PBS. Loading and washing were performed in 0.1 M $NaHCO_3$, 0.5 M NaCl, 1 mM EDTA, pH 8.0, and elution in the 0.1 M glycine, 0.5 M NaCl, 10% dioxane, pH 2.8. The binding-fraction (Lp-AI particles) was collected at 4 °C as a single 5 ml fraction on a Frac 900 fraction collector (GE Healthcare). After elution, 0.5 M Tris–HCl, pH 8.0

was added to neutralize the pH and prevent protein degradation. The purified Lp-AI fraction was concentrated to approximately 200 μl at 3000 rpm for 35–40 min and 4 °C, using Amicon Ultra 10 kDa filters (Millipore). Total protein concentration was calculated by RC/DC Protein assay (BioRad), a Lowry based commercial method, with albumin as the standard.

2.3. SDS-PAGE and western blotting for ApoA-I

In order to determine the yield of purified Lp-AI particles, samples were separated by conventional 10% SDS-PAGE and then transferred to nitrocellulose membranes. Membranes were probed with goat anti-human ApoA-I antibody (Santa Cruz Biotechnology) and signal was developed with the Amersham ECL Plus Western Blotting Detection system (GE Healthcare), according to the manufacturer's instructions.

2.4. In-gel tryptic digestion

Protein extracts were suspended in Laemmli sample buffer and separated by SDS-PAGE (1.5 mm-thick gel, 4% stacking, 10% resolving). The run was stopped as soon as the front entered 3 mm into the resolving gel, so that the whole proteome became concentrated in the stacking-gel/resolving-gel interface, as described [14,15]. Protein bands were visualized by Coomassie staining, excised, and cut into cubes (2 × 2 mm). Protein extracts were digested using a robust protocol [14], by overnight incubation at 37 °C with 60 ng/μl trypsin at a 5:1 protein:trypsin (w/w) ratio in 50 mM ammonium bicarbonate, pH 8.8 containing 10% (v/v) ACN and 0.01% (w/v) Cymal-5. The resulting peptides were desalted onto C18 OASIS cartridges (Waters) and dried-down.

2.5. ^{18}O labeling

Two groups of HDL samples were used for quantitative $^{16}O/^{18}O$ analysis. One group (a ^{16}O-labeled 400 μg pool from 7 patients before PTCA versus an ^{18}O-labeled 400 μg pool from the same patients after PTCA) was used to compare the HDL proteome as an alternative hypothesis. The second group (a ^{16}O-labeled 400 μg pool from 4 patients versus an ^{18}O-labeled 400 μg pool from another 4 patients, all before PTCA) was used to test the null hypothesis. ^{18}O labeling was performed as described [14]. Briefly, dried peptides from the pools were subjected to differential $^{16}O/^{18}O$-labeling in 100 mM ammonium acetate, pH 6.0, 20% (v/v) ACN, at a 1:200 (v:w) immobilized trypsin/protein ratio. After labeling, trypsin beads were removed using a physical filter (Wizard minicolumns, Promega) and by adding 1 mM TLCK to the filtrate and incubating for 1 h at 37 °C. The two labeled samples were mixed, diluted to 2% (v/v) ACN with 1 M ammonium formate pH 3.0, desalted onto C18 Oasis cartridges (Waters) using 50% (v/v) ACN in 5 mM ammonium formate pH 3.0 as the elution solution, and dried down.

2.6. iTRAQ labeling

Another two sample groups were used for quantitative iTRAQ analysis. Each group contained four 100 μg protein extracts obtained from two patients before and after PTCA. Samples obtained before PTCA were labeled with 114 and 116 tags and samples from the same patients after PTCA were labeled with 115 and 117 tags. For iTRAQ labeling, dried peptides were taken up in 30 μl iTRAQ dissolution buffer (Applied Biosystems) and labeled with 70 μl of the corresponding iTRAQ reagent in 70% (v/v) ethanol and 180 mM triethylammoniumbicarbonate (TEAB), pH 8.53, for 1 h at room temperature. After this, 200 μl 0.1% (v/v) formic acid was added and samples were brought to dryness to completely stop the labeling reaction. The pre- and post-PTCA samples from each pair of patients were resuspended in 200 μl 0.1% (v/v) formic acid and mixed. The mixture was dried down, redissolved in 12 ml of 5 mM ammonium formate, pH 3.0, cleaned up with MCX Oasis cartridges (Waters, Milford; Massachusetts, USA) using as elution solution 1 M ammonium formate, pH 3.0 containing 25% (v/v) ACN, and dried down. The peptide pools were resuspended in 0.5 ml 0.1% (v/v) TFA, desalted onto C18 Oasis cartridges using 50% (v/v) ACN in 0.1% (v/v) TFA as elution solution, and dried down.

2.7. IEF peptide fractionation

$^{16}O/^{18}O$-labeledpeptide pools were taken up in focusing buffer (GE Healthcare), loaded onto 24 wells over a 24 cm-long Immobiline DryStrip, pH 3-10 (GE Healthcare) [14], and separated by IEF on a 3100 OFFgel Fractionator (Agilent Technology, Santa Clara, CA), using the manufacturer's recommended method. Peptide fractions were desalted using OMIX C18 tips using 50% (v/v) ACN in 5 mM ammonium formate, pH 3.0 as elution solution, and dried down.

2.8. SCX peptide fractionation

iTRAQ-labeled peptide pools were separated into 8 fractions using MCX Oasis cartridges, using elution buffers of 0.5, 1.0, 1.5, and 2.0 M ammonium formate, pH 3.0, containing 25% (v/v) ACN, and recovering 2 eluted fractions for each salt concentration. Peptide fractions were desalted using OMIX C18 tips, as described above, and dried down.

2.9. LC–MS/MS analysis and peptide identification

$^{16}O/^{18}O$-labeled samples were analyzed by LC–MS/MS, using a Surveyor LC system coupled to a linear ion trap LTQ (Thermo-Finnigan, San Jose, CA, USA). Peptides were concentrated and desalted onto an RP precolumn (0.32 × 30 mm, BioBasic C18, Thermo Electron) and eluted on line onto an analytical RP column (0.18 × 150 mm, BioBasic C18, Thermo Electron) operating at 2 μl/min and using the following gradient: 5% B for 15 min, 5–14% B in 15 min, 14–30% B in 155 min, 30–95% B in 7 min, and 95% B for 3 min (solvent A: 0.1% formic acid; solvent B: 0.1% formic acid, 80% CH3CN) [16–18]. The LTQ was operated in data-dependent ZoomScan- and MS/MS-switching mode [16]. iTRAQ-labeled samples were analyzed by LTQ in the PQD scanning mode, operated in data-dependent MS/MS on the 15 most-intense precursors detected in a full scan from 400 to 1600 amu. Zoom target and PQD parameters, number of microscans, normalized collision energy, dynamic exclusion parameters were as previously described [16–18]. LC conditions were as described [18]. Peptides were identified by searching a joint human database (Uniprot release 1512 Jun 2010) supplemented with porcine trypsin, and using the SEQUEST algorithm

(Bioworks 3.2 package, Thermo Finnigan), as previously described [19]. For $^{16}O/^{18}O$-labeled samples, variable modifications (methionine oxidation, lysine and arginine modification of + 4 Da) and fixed modifications (cysteine carboxamidomethylation) were used. For iTRAQ, variable modifications (methionine oxidation) and fixed modifications (cysteine carboxamidomethylation, lysine and N-terminal modification of +144.1020 Da) were allowed. The same collections of MS/MS spectra were also searched against inverted databases constructed from the same target databases. SEQUEST results were analyzed using the probability ratio method [20], taking into account isoelectric points of peptides to improve peptide identification [14]. False discovery rates (FDR) of peptide identifications were calculated using the refined method [21] from the results of searching the inverted databases.

2.10. Peptide quantification and statistical analysis

Peptides were quantified from Zoom Scan and PQD MS/MS spectra and ^{18}O labeling efficiencies calculated for peptides identified with an FDR <5% using QuiXoT, a program written in C# in our laboratory [14]. Statistical analysis of the data was done on the basis of a novel random-effects model that includes four independent sources of variance: at the spectrum-fitting, scan, peptide and protein levels [18,22]. In each scan s, the $\log_2$-ratio of the concentrations of peptide p derived from protein q in non labeled (A) and labeled (B) samples is expressed as $x_{qps} = \log_2(A/B)$. The overall $\log_2$-ratio value associated with each peptide, x_{qp}, is calculated as a weighted average of the scans used to quantify the peptide, and the value associated with each protein, x_q, is similarly the weighted average of its peptides. A grand mean, x, is calculated as a weighted average of the protein values. In addition, the statistical weight associated with the scan, w_{qps}, is calculated from the spectrum fitting and the scan variance, σ_S^2 [18]. The statistical weight associated with each peptide, w_{qp}, is calculated from the corresponding scan weights and the peptide variance, σ_P^2, and that of each protein, w_q, is calculated from the corresponding peptide weights and the protein variance, σ_Q^2 (the statistical weights are the inverses of variances). A standardized variable z_i is defined at each level as the mean-corrected $\log_2$-ratio expressed in units of standard deviation; this variable allows the representation of all the data, regardless of its actual variance, in a common normal distribution. Details of the statistical model and the algorithm used to calculate the variances at the scan, peptide and protein levels can be found in our previous works [18,22]. In all cases, the proportion of partially digested peptides was lower than 10%, an efficiency of digestion similar to that of previously analyzed proteomes [14]. Labeling efficiencies of each peptide were carefully controlled using a proposed method [17], and were similar to those typically obtained by using the robust protocol [14]. Similarly, the number of outliers at spectrum and peptide levels was negligible. All these parameters were carefully controlled and were similar in all experiments performed.

2.11. Integration of data and system biology analysis

Integration of quantitative data obtained for each one of proteins across different experiments was done as we have described [22]. For each protein q, an averaged protein $\log_2$-ratio x'_q is obtained as a weighted average of the mean-corrected values of q in all the experiments. The statistical weight associated to the averaged protein values, w'_q, is calculated from the corresponding protein weights and the average protein variance, σ'^2_Q. To perform system biology analysis, quantified proteins were grouped into all possible functional categories according to GO (http://www.geneontology.org), Reactome (http://www.reactome.org/) and KEGG annotations (http://www.genome.jp/kegg/). Similarly, an averaged $\log_2$-ratio x''_c was calculated for each category as a weighted average of the proteins that belong to the category. A standardized variable z_i is defined at each integration level, to describe the distribution of proteins around the protein averages, the distribution of proteins, the distribution of proteins around the category averages and the distribution of category values, as explained above. The exact mathematical details are described in the Supplementary Information. Categories with at least four proteins were subjected to hierarchical cluster analysis, using Ward algorithm with Euclidian distance as implemented on R Commander.

3. Results

3.1. Isolation and characterization of HDL-associated proteins

In this study, HDL particles were isolated from human plasma by immunoaffinity chromatography using an antibody against human ApoA-I. This method allowed fast and effective particle purification directly from plasma. We first analyzed the specificity of the human anti-ApoA-I antibody by comparing the eluted fractions obtained when identical plasma samples were fractioned on an ApoA-I antibody-coated column and a blank column not containing the antibody. SDS-PAGE and western-blot confirmed that only the eluted fraction from the antibody-coated column was specifically enriched in ApoA-I (Supplementary Fig. 1S). A preliminary mass spectrometry analysis confirmed the enrichment in ApoA-I in the antibody-bound fraction and also demonstrated a decrease in the proportion of serum albumin (Supplementary Fig. 2S). This analysis also revealed the presence of paraxonase/arylesterase 1, Apo-D, complement factor H, the α chain of fibrinogen and the C region of the immunoglobulin α-1 chain (Supplementary Fig. 2S). All these proteins are known HDL constituents [9,23,24] and were not found in the fraction eluted from the blank column. We therefore concluded that the affinity-purified HDL preparation was reasonably enriched in HDL particles according to their protein cargo, and was hence suitable for high-throughput proteomics analysis. We should note that, although the vast majority of human HDL particles are known to contain ApoA-I, a small proportion does not. The particles isolated by this method are therefore not expected to comprise all plasma HDL particles, but only the dominant Lp-AI (ApoA-I containing) fraction.

The protein composition of the HDL preparation was analyzed in detail using six high-throughput quantitative proteomics experiments using stable isotope labeling (see Materials and methods). To increase stringency, we filtered out peptides identified with a false discovery rate higher than

0.01, and only considered proteins identified in at least two experiments, producing a refined list of HDL-associated proteins (Supplementary Excel Table 2 and Supplementary Fig. 3S). From a total of 225 identified proteins, 66 (30% of the total) have been previously identified in HDL isolated by techniques such as ultracentrifugation [23–31] and gel filtration [8] and affinity [9,31] chromatographies. These included 15 apolipoproteins, fibrinogen α, β and γ chains, α1-antitrypsin, α2-HS-glycoprotein, antithrombin-III, complement C3, kininogen-1, serum amyloid A, serum paraoxonase/arylesterase 1 and thrombospondin-1. Another 47 proteins (20% of the total) were immunoglobulins, four of which have been recently described as forming part of HDL [24]. The remaining 111 (50%) proteins have not been previously described in HDL preparations, and included ApoA-V, proteins of the ATP-binding cassette family, many complement component proteins, peroxiredoxins, kinases, cadherins, and coagulation factors.

3.2. Analysis of pooled samples to detect angioplasty-induced alterations in the HDL proteome

HDL samples, each containing 400 μg total protein, were obtained by pooling HDL preparations from 7 patients before and after PTCA. After protein extraction, digestion, labeling, fractionation and LC–MS/MS analysis, 8547 MS/MS spectra were assigned to peptide sequences, corresponding to 894 unique peptides and 324 proteins. A total of 4947 Zoom scan spectra could be quantified, corresponding to 427 unique peptides and 134 proteins. The quantitative data were analyzed by a statistical model that controls the variance at the spectrum, peptide and protein levels [18,22]. The distribution of the standardized variables at both the spectrum and peptide levels [18] followed the trend expected for the null hypothesis (Fig. 1A and B, black points and red lines, respectively), and the calculated variances at these two levels were similar to those obtained with other proteomes [14,18] (Fig. 2; see also Supplementary Excel Table 3). Consistently, partial protein digestion and methionine oxidation, parameters known to produce quantification artifacts [14,18], were not found to introduce any bias to the peptide quantification (not shown). These essential quality tests indicated that MS analysis and peptide preparation and labeling were within the expected standards.

The standardized protein distribution also followed the null-hypothesis trend (Fig. 1C). However, the distribution of protein quantifications around the experimental grand mean was much larger than expected (Fig. 1D), and the calculated variance at the protein level (0.23; Fig. 2, light blue bar) was about two orders of magnitude higher than the value obtained in other proteomes we have analyzed to date using the same statistical model, including immortalized and primary human cell lines and animal tissues [14,18] (Supplementary Excel Table 3). Although human samples are known to display high inter-individual variability and therefore the comparison with these other models should be considered with caution, determining the actual source of this apparently high protein variability was essential for judging whether the dynamic behavior of the HDL proteome could be studied on a routine basis. Therefore, in a second quantitative experiment, which we call the "null hypothesis" experiment, we eliminated the

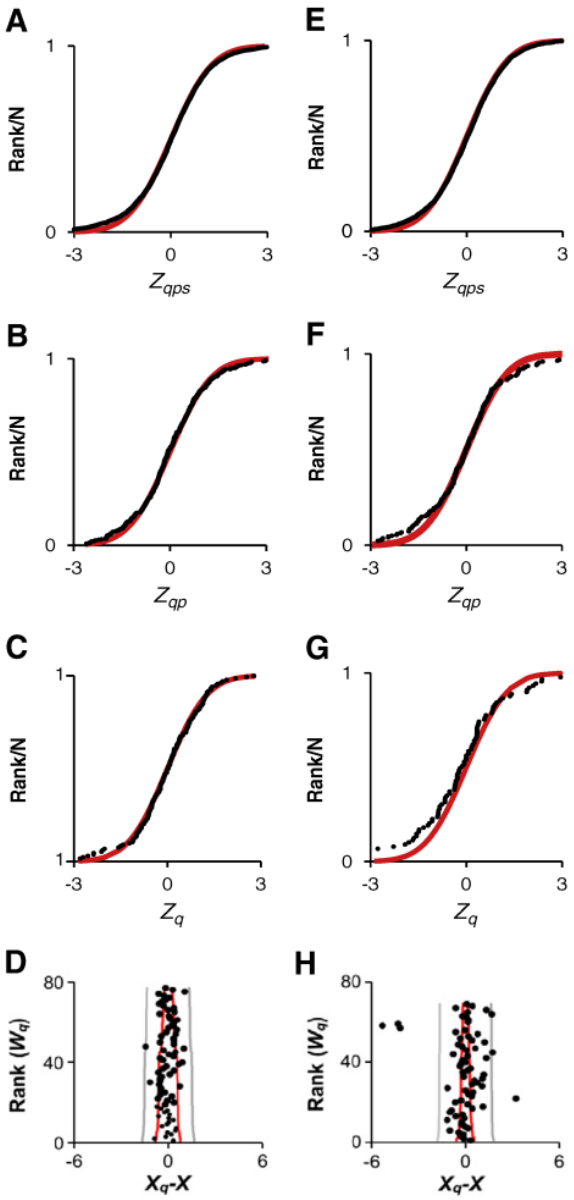


Fig. 1 – Quantitative ^{18}O-labeling analysis of the effect of PTCA on HDL samples pooled from several individuals. A–D, Analysis of two pooled samples from CAD patients before and after PTCA treatment (seven patients each pool). E–H, Analysis of two pooled samples from CAD patients not treated by PTCA (four patients each pool). The graphs are the cumulative distributions of the standardized variables at the scan level, z_{qps} (A, E), at the peptide level, z_{qp} (B, F) and at the protein level, z_q (C, G), showing the agreement between the experimental data (black points) and the theoretical trends (red lines). D and H show the weight distributions of protein quantifications (log_2-ratios) around the grand mean. The gray lines indicate confidence intervals corresponding to 5% FDR. For illustrative purposes we have also drawn the 5% FDR confidence intervals (red lines) that would have been obtained if the protein variance of the experiment were 0.005.

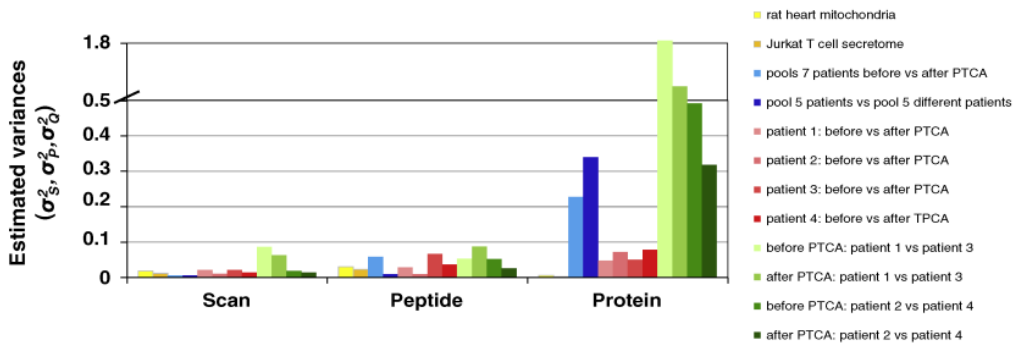


Fig. 2 – Comparison of variances estimated at the scan, peptide, and protein levels from the quantitative analysis of HDL proteomes. The variances at scan, peptide, and protein levels (σ^2_S, σ^2_P, σ^2_Q, respectively) from all experiments performed are represented. For illustrative purposes the variances obtained in previous experiments, performed in rat heart mitochondria and Jurkat T cell secretome [14] are also plotted.

effect of angioplasty by comparing protein abundance in two pooled HDL samples from different CAD patients not subjected to this procedure. All technical quality parameters, including all standardized distributions (Fig. 1E–G) and variances at the spectrum and peptide levels (Fig. 2 and Supplementary Excel Table 3), were similar to the previous

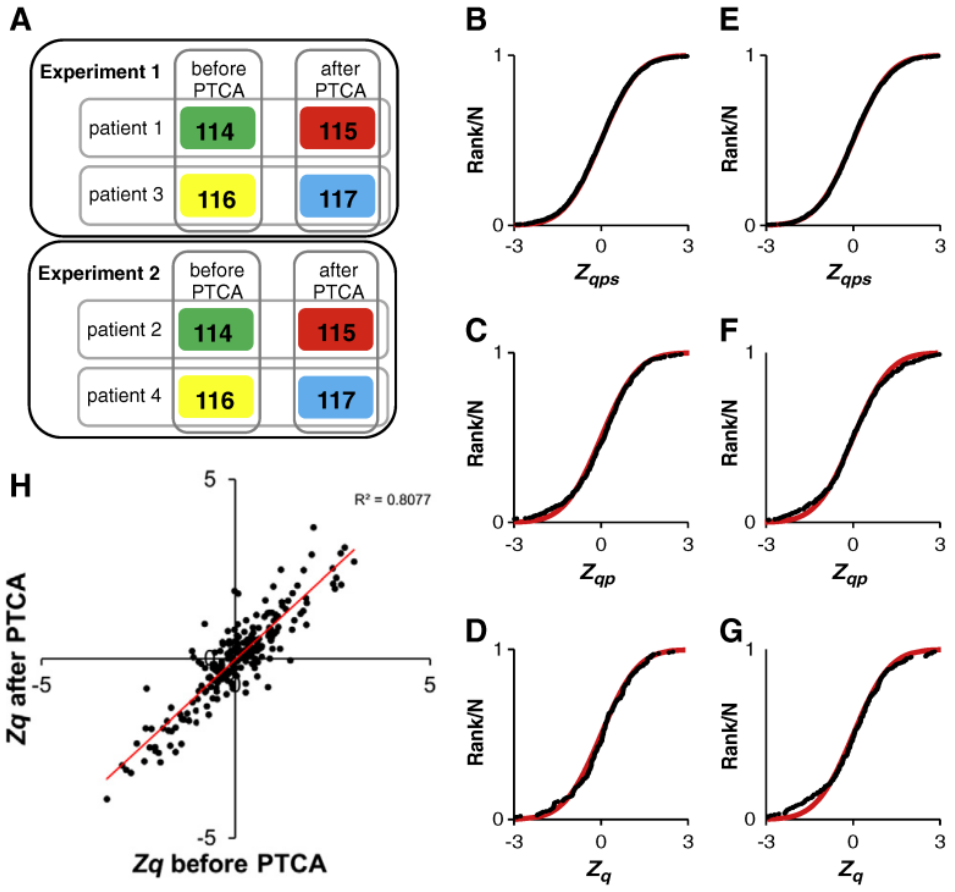


Fig. 3 – Quantitative proteomics analysis by iTRAQ-labeling of the effect of PTCA on HDL in individual CAD patients. A, Scheme of the two quantitative experiments, indicating the iTRAQ reporters used to label peptides derived from each HDL preparation and the pairwise comparative analysis performed. B–G, Cumulative distributions of the standardized variables at the scan, peptide and protein levels; plots are as in Fig. 1. B–D, Comparative analysis of HDL samples obtained from patient 3 before and after PTCA. E–G, Comparative analysis of HDL samples obtained from patient 1 and patient 3, both before PTCA. H, Correlation of HDL protein abundance in patient 1 and patient 3 before and after PTCA.

experiment. However, the protein variance (0.34, Fig. 2 dark blue bar) remained two orders of magnitude above the standards of our quantification method (Supplementary Excel Table 3). These results confirmed that the observed variability in the HDL proteome was not induced by angioplasty, but was present in the samples that were subjected to quantitative analysis.

3.3. *Multiplexed analysis of biological variability and the effect of angioplasty on individual HDL proteomes*

Since the large variability in HDL protein abundance could be due to large inter-individual variability or poor reproducibility of HDL preparations, we devised a multiplexed strategy to separately control these two factors. HDL preparations from four individuals were obtained before and after PTCA treatment and the eight samples were compared in two independent 4-plex iTRAQ experiments. The digested samples were labeled with iTRAQ reagents (Fig. 3A), with the idea that the variances calculated by the eight indicated pairwise comparisons would give separate information about the corresponding error sources. The labeled pre- and post-PTCA peptides from each sample were mixed, SCX-separated into 8 fractions, and analyzed by linear ion trap mass spectrometry using PQD fragmentation.

The standardized distributions at the spectrum, peptide and protein levels in HDL samples from patient 3 before and after PTCA (Fig. 3B–D), and from patient 1 and patient 3 before PTCA (Fig. 3E–G), demonstrated that the quantitative iTRAQ results were also in good agreement with the predictions of the statistical model and therefore that the model was adequate for analysis of this kind of data. This was confirmed by analysis of sigmoidal distributions obtained in the remaining pairwise comparisons (Supplementary Fig. 4S). Moreover, variances at the spectrum and peptide levels were similar in all the eight pairwise comparisons, and were very like those obtained by $^{16}O/^{18}O$ labeling in the two previous experiments and to other results obtained previously in our laboratory [14,18] (Fig. 2 and Supplementary Excel Table 3), indicating that all technical parameters related to iTRAQ quantification passed the quality tests. When HDL proteomes from different patients were compared both before and after PTCA, the variances at protein level had the same large values observed in the previous experiments (Fig. 2, green bars and Supplementary Excel Table 3). In clear contrast, the protein variances calculated for comparisons of samples prepared before and after PTCA in the same individual were below 0.1 in all four cases, similar to values found in other proteomes [14,18] (Fig. 2, red bars and Supplementary Excel Table 3). These results clearly demonstrated that the high variability found in the initial experiments was not due to the HDL isolation

Table 1 – List of most variable proteins in HDL preparations from CAD patients not subjected to PTCA.

		Corrected log_2–ratio (X_q) [a]			Standardized log_2–ratio (Z_q) [b]		
Acc number [c]	Protein name [d]	NH [e]	P1 vs P3 [f]	P2 vs P4 [f]	NH [e]	P1 vs P3 [f]	P2 vs P4 [f]
P02671	Fibrinogen alpha chain	−4.46	4.29	1.24	−17.31	13.90	4.68
P02679	Fibrinogen gamma chain	−4.62	3.55	0.96	−17.73	12.16	3.65
P02675	Fibrinogen beta chain		2.35	0.81		8.08	3.13
P01008	Antithrombin–III	1.73	1.11		7.32	2.75	
P00734	Prothrombin	1.78	1.98		7.32	5.17	
O14791	Apolipoprotein L1		0.71	1.65		2.27	5.34
P02655	Apolipoprotein C–II		1.53	0.93		4.97	3.26
P02647	Apolipoprotein A–I	0.48	1.13		2.35	4.13	
P35542	Serum amyloid A–4 protein	0.66	1.30		2.98	3.91	
P01842	Ig lambda chain C regions	0.83	−0.93		3.68	−3.01	
P02790	Hemopexin		1.12	−0.77		3.42	−2.42
P01023	Alpha–2–macroglobulin	−0.94	0.81	−0.74	−3.28	3.05	−2.84
P68871	Hemoglobin subunit beta		0.81	−0.74		2.63	−2.55
P04114	Apolipoprotein B–100		−0.71	0.54		−2.82	2.13
P00738	Haptoglobin	−0.99		−0.66	−3.44		−2.19
P02751	Fibronectin	−1.66	−0.80		−6.03	−2.62	

[a] Log2-ratios.
[b] Standardized normal values obtained by dividing protein log_2-ratios by their variance. Proteins are filtered by $|z_q| > 2.0$.
[c] Uniprot accession number.
[d] Proteins quantified in two or more experiments and with at least two single peptides.
[e] Null hypothesis experiment.
[f] Individual patients before PTCA from iTRAQ experiments.

protocol, which introduced only a very low technical variability in the proportion of HDL proteins and therefore can be considered appropriate for quantitative analysis. Our results rather demonstrate that HDL protein composition varied significantly from one patient to another: in other words, that the HDL proteome has a very large biological variability, being highly patient-specific. However, plotting the standardized protein $\log_2$-ratios calculated for comparison of patient 1 with patient 3 before PTCA against those calculated for the same comparison after PTCA revealed a strong linear correlation (Fig. 3H). The good agreement in the magnitude and sign of protein quantifications can also be appreciated in Supplementary Excel Table 4. These results indicate that the high HDL protein variability among individuals was exquisitely maintained before and after PTCA, and therefore that the analysis of the effect of PTCA on HDL proteome could be performed in samples isolated from the same individuals.

3.4. Analysis of protein variability in HDL preparations from different individuals

The protein variances estimated from the comparisons of HDL preparations before and after PTCA in the same individuals can be considered as conservative estimates of the technical protein variance associated with the process of HDL preparation. We therefore calculated the average protein variance in the four experiments (0.062) and used it to estimate the null-hypothesis distribution of standardized $\log_2$-ratios of proteins. We used this distribution as a new reference to determine statistically significant protein abundance changes in the HDL preparations between each pair of individuals before PTCA and also the protein abundance changes between the two pooled HDL samples from different patients not subjected to PTCA. Considering only the proteins that changed their abundance in at least two comparisons and that were quantified with two or more unique peptides, we obtained a curated list of 16 proteins that showed significant abundance changes. This list comprises the HDL proteins with the highest biological variability among different individuals (Table 1 and Supplementary Excel Table 5). Eleven of these, ApoA-I, Apo-LI, ApoC-II, serum amyloid A-4 protein, Igλ chain C, hemopexin, α-2-macroglobulin, β-subunit of hemoglobin, apolipoprotein B-100, haptoglobin and fibronectin, were among the most abundant proteins identified in the HDL proteome. Other proteins showing high biological variability were the α-, β- and γ-chains of fibrinogen, antithrombin-III, and prothrombin.

3.5. Analysis of intra-individual alterations in HDL protein composition induced by PTCA

Most protein abundance changes induced by PTCA were consistently reproduced in the four individuals (Table 2). To determine the global pattern of HDL proteins showing statistically significant abundance changes as a consequence of PTCA, we integrated the results obtained from the four patients using a statistical model recently proposed in our laboratory [22]. This model allows comparison and coherent integration of results obtained from different analyses according to error propagation theory, thereby increasing the statistical power of protein quantification and also allowing the detection of outliers in the different experiments. After experiment integration, quantitative data from a total of 200 proteins was obtained (Supplementary Excel Table 6). The distribution of the standardized variable describing the variability between different experiments within the same protein was very close to the null hypothesis (Fig. 4A), demonstrating that quantifications from all experiments were generally reproducible and therefore can be integrated. A subset of proteins, however, had $\log_2$-ratios significantly different from the integrated value (Supplementary Excel Table 7); these discrepancies were presumably due to the high interindividual variability of the HDL proteome. Patient 4 appeared to be the most discrepant, showing protein abundance changes in fibrinogen subunits and other proteins opposite to those in the other patients. Some proteins from patient 2 also deviated from the average; of these, ApoL-I, platelet basic protein and Igλ chain C also deviated in patient 4. For statistical consistency, these protein outliers were eliminated from the analysis.

The distribution of standardized protein $\log_2$-ratios after integration of the four experiments also followed the expected trend for the null hypothesis (Fig. 4B). Five of these proteins were found to be significantly upregulated after PTCA treatment, while another 4 proteins were downregulated (Fig. 4C and Table 2). A group of apolipoproteins, including ApoC-IV, ApoC-II, and ApoA-IV, were increased after PTCA treatment. Cytoplasmic actin 1 and liver-specific fibrinogen-like protein 1 were also significantly increased. One of the most markedly decreased proteins was antithrombin-III, and other downregulated proteins included Igλ-1 chain C region (IgLC1), Ig heavy chain variable region, and annexin A1.

3.6. Systems biology analysis of angioplasty-induced alteration in the HDL proteome

To obtain a clearer picture of PTCA-induced alterations in the HDL proteome, we performed an integrative quantitative analysis of the behavior of ontological categories, extending the same framework proposed previously [22] (Supplementary Information). All quantified proteins were classified into all possible functional GO annotations and KEGG pathways, producing a collection of averaged category values. The distribution of protein quantifications within each of the categories followed a normal distribution with unit variance, and only 1.67% of proteins were outliers in their categories (Fig. 4D), suggesting a predominantly coordinated behavior, with most proteins acting in the same way as the other proteins in the same category. We also analyzed the distribution of the standardized variable describing the inter-category variability (Fig. 4E); interestingly, 34 out of 494 categories with at least 2 proteins were significantly increased at a 5% FDR, while no categories where decreased. This finding suggests that PTCA increased the HDL protein cargo predominantly in specific functional categories. The quantitative data for all the altered categories are shown in Supplementary Excel Table 8.

These 34 functional categories have highly overlapping protein compositions (Supplementary Excel Table 9); we therefore performed a clustering analysis, using only

categories with more than 4 proteins, from which two main groups emerged that contained the majority of proteins. One group was linked to biological processes related to apolipoproteins (Supplementary Fig. 5S, blue cells). The second group contained proteins related to ATP and nucleotide binding, focal adhesion and tight junctions (Supplementary Fig. 5S, orange cells), and also intracellular proteins such as actins, myosins, kinases, phosphatases and other regulatory factors and several transmembrane, ATP-binding and transporter proteins. The global pattern of protein abundance changes in these two main groups is shown in Fig. 4F, where the increase in abundance of these proteins as a consequence of PTCA is clearly observed.

4. Discussion

To the best of our knowledge, this is the first study published to date that studies the effect of an angioplasty treatment in the protein composition of circulating HDL. In addition, 111 novel proteins have been identified associated to CAD HDL particle, making this study the deepest proteomics analysis of

Table 2 – Protein abundance changes in HDL proteome of CAD patients subjected to PTCA treatment. List proteins are quantified in at least two patients and include proteins that present a statistically significant change in abundance at 10% FDRq′ criterion, which are numbered and indicated in Fig. 4. The magnitudes of the abundance change and of the standardized variable are shaded according to the color scale at the bottom.

			Corrected log_2–ratio (X_q) [a]				Standardized log_2–ratio (Z_q) [b]				Integrated analysis		
			EXPT 1		EXPT 2		EXPT 1		EXPT 2				
No	Acc number [c]	Protein name [c]	P1 [e]	P3 [e]	P2 [e]	P4 [e]	P1 [e]	P3 [e]	P2 [e]	P4 [e]	X_q' [a]	Z_q' [b]	FDR_q' [f]
1	P55056	Apolipoprotein C–IV	−0.43	−0.79	−0.45	−1.27	−1.39	−1.61	−1.50	−3.92	−0.71	−4.19	0.003
2	P02655	Apolipoprotein C–II	−0.19	0.02	−0.90	−1.06	−0.68	0.06	−3.31	−3.48	−0.50	−3.59	0.017
3	P60709	Actin, cytoplasmic 1	−1.07	−0.29			−3.36	−0.94			−0.67	−3.01	0.087
4	P06727	Apolipoprotein A–IV	−0.30	−1.27	0.83	−0.58	−1.06	−4.22	1.91	−1.60	−0.49	−2.95	0.079
5	Q08830	Fibrinogen–like protein 1	−0.24	−1.19	0.11	−0.48	−0.82	−3.81	0.30	−1.27	−0.48	−2.90	0.075
	Q9BXM7	Serine/threonine–protein kinase PINK1, mitochondrial	−0.03	−0.93			−0.07	−2.78			−0.64	−2.31	0.345
	Q9BRS2	Serine/threonine–protein kinase RIO1	−0.79		0.65	−1.20	−2.34		1.80	−3.32	−0.46	−2.27	0.329
	Q15195	Plasminogen–related protein A	−0.46	−0.51	−0.10	−0.14	−1.50	−1.77	−0.29	−0.40	−0.34	−2.09	0.455
	Q9NP78	ATP–binding cassette sub–family B member 9			−0.10	−0.94			−0.29	−2.60	−0.52	−2.03	0.470
	P02656	Apolipoprotein C–III	−0.02	−0.55	−0.25	−0.35	−0.09	−1.89	−0.85	−1.11	−0.29	−1.96	0.507
	Q92954	Proteoglycan 4	−0.35	−0.41			−1.09	−1.62			−0.39	−1.95	0.468
	P07225	Vitamin K–dependent protein S		−0.84	−0.12	−0.37		−2.11	−0.36	−0.99	−0.41	−1.92	0.460
	P02649	Apolipoprotein E	0.02	−0.69	−0.12	0.16	0.07	−2.91	−0.48	0.55	−0.20	−1.56	0.909
	Q8WW33	Gametocyte–specific factor 1			−0.59	−0.21			−1.59	−0.57	−0.41	−1.53	0.901
	P08519	Apolipoprotein(a)	−0.72	0.26	−0.37		−2.54	0.96	−1.18		−0.25	−1.53	0.849
	O95445	Apolipoprotein M	−0.21	0.02	−0.70	−0.03	−0.76	0.08	−2.37	−0.08	−0.22	−1.51	0.816
	P02751	Fibronectin	−0.64	0.04	−0.07	−0.41	−2.16	0.15	−0.23	−0.99	−0.23	−1.51	0.776
	Q9HB19	Pleckstrin homology domain–containing family A member 2	0.91	0.46			1.48	0.73			0.69	1.57	0.619
	P02671	Fibrinogen alpha chain	0.55	−0.31	0.54		2.05	−1.32	2.18		0.23	1.58	0.636
	P01042	Kininogen–1	0.14	0.46			0.49	1.81			0.32	1.68	0.554
	P01825	Ig heavy chain V–II region NEWM	1.00	0.73			1.19	1.27			0.82	1.72	0.541
	B3F343	HHV8 K1 protein (Fragment)	1.30	−0.11			2.17	−0.12			0.87	1.75	0.538
	P04114	Apolipoprotein B–100	0.53	−0.08	0.33	0.23	1.98	−0.37	1.43	0.80	0.23	1.82	0.489
	P54259	Atrophin–1	1.04	0.73			1.63	0.95			0.92	1.86	0.489
	P62988	Ubiquitin	1.38	0.56			1.77	1.11			0.80	1.89	0.491
	A0M8Q6	Ig lambda–7 chain C region	0.32	0.53			0.98	1.71			0.43	1.92	0.500
	P00734	Prothrombin	0.97	−0.07			3.12	−0.24			0.45	2.04	0.419
	P02679	Fibrinogen gamma chain	0.48	0.04	0.45		1.74	0.15	1.87		0.31	2.15	0.353
	Q6N095	Putative uncharacterized protein DKFZp686K03196	0.41	0.55			1.37	1.79			0.48	2.23	0.321
	C0JYY2	Apolipoprotein B (Including Ag(X) antigen)	0.51	−0.03	0.52	0.50	1.79	−0.12	1.89	1.56	0.35	2.48	0.188
	Q5NV83	V3–3 protein (Fragment)	0.40	1.84			0.98	3.15			0.87	2.60	0.157
6	Q0ZCF9	Immunglobulin heavy chain variable region (Fragment)	0.95	0.32			3.06	0.93			0.66	2.88	0.099
7	P04083	Annexin A1	0.12	1.91			0.14	3.41			1.36	2.92	0.116
8	P01842	Ig lambda chain C regions	0.01	0.40	1.96		0.01	1.58	5.66		0.62	3.73	0.010
9	P01008	Antithrombin–III	1.16	0.07	0.09	1.50	3.75	0.22	0.25	4.14	0.72	4.24	0.002

Color scale at the bottom.

−1	−0.5	0	0.5	1	X_q
−5	−2.5	0	2.5	5	Z_q

[a] Log_2-ratios in each comparative experiment (x_q) and in the integrated statistical analysis (x'_q).
[b] Standardized normal values calculated by dividing protein log_2-ratios by their variance obtained in each comparative experiment (z_q) and in the integrated statistical analysis (z'_q). Proteins are filtered by $|z'_q| > 1.5$.
[c] Uniprot accession number.
[d] Proteins quantified in two or more experiments and with at least two single peptides.
[e] CAD patients subjected to PTCA treatment.
[f] False discovery rate at protein level in the integrated statistical analysis.

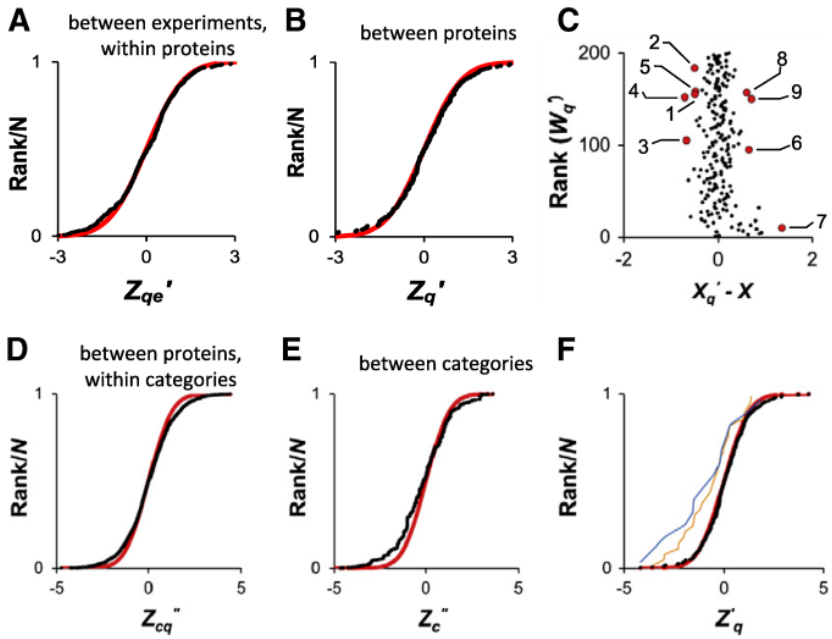


Fig. 4 – Integrative analysis of HDL protein abundance changes produced by PTCA in four CAD patients and functional analysis of HDL proteome alterations. Data from the quantitative experiments studying the effect of PTCA on the HDL proteome in four CAD patients were integrated to obtain averaged protein values as described [22]; the statistical results of the integration are shown in A–C. A, Cumulative distribution of the standardized variable at the protein level (z'_{qe}), which describes the variability of the quantifications for the same proteins across the four experiments; deviations of the observed data (black points) from the null hypothesis (red curve) indicate the presence of outliers, where quantifications of a protein in one experiment differ significantly from the quantifications of the same protein in the rest of experiments. B, Cumulative distribution of the standardized variable at the protein level (z'_q), which measures averaged protein abundance changes produced by PTCA in the four individuals; the deviations from the null-hypothesis indicate significant protein abundance changes. C, Weight distributions of averaged protein values (log_2-ratios) around the grand mean; numbered dots indicate significant increases (leftwards) or decreases (rightwards) in protein abundance quantified in at least two patients, as described in Table 2 and in Supplementary Excel File T7. The proteins were classified into all possible functional GO annotations and KEGG pathways, and category averages were calculated as described [22]. D, Cumulative distribution of the standardized variable at the protein level within each category (z''_{cq}), which describes the variability of the quantifications of proteins belonging to each category. E, Cumulative distribution of the standardized variable at the category level, which measures averaged category abundance changes produced by PTCA. F, As B, but showing the coordinated increase of proteins belonging to the two main protein groups obtained by clustering analysis: the group of apolipoproteins (blue line) and the group of intracellular and transmembrane proteins (orange line).

HDL proteome. Biological process according to GO mainly showed that 33 proteins were implicated in primary metabolism, 20 proteins were associated to immune system, 17 proteins were involved in transport, 15 proteins were included in proteolysis process, 15 proteins were of cellular communication, and 6 proteins were implicated in extracellular matrix. Interestingly, 9 complement components non-previously described associated to HDL have been identified in our study. These results validate the previous role of HDL in complement regulation and proteolysis [9] and suggest that HDL could play a key role in modulation of extracellular matrix degradation and participate in reverse transport of proteins from the atheroma plaque.

The validity of our results relies mainly on the method used to isolate the HDL particles. Classically, HDL and other lipoproteins have been isolated by density gradient ultracentrifugation [32], which has the potential disadvantage that the protein composition of HDL can be altered by the violent shear conditions and high salt concentration needed. To protect the weak bonds that link proteins to the HDL particle, we instead purified HDL by affinity chromatography using a commercially available antibody against ApoA-I, the most abundant HDL protein component. Isolation by this procedure involves non-aggressive physiological salt and shear conditions and is relatively fast. Moreover, the method avoids contamination of HDL samples with high abundance plasma proteins and lipoproteins that occur when isolation is performed by gel filtration chromatography [8]. Our method has, however, the inconvenience that only ApoA-I-containing particles (LpA-I) are isolated. Although these constitute the vast majority of HDL particles [33], it is important to note here that the conclusions of this study cannot be extrapolated to the small proportion of non-ApoA-I containing HDL particles.

Alterations in the protein composition of HDL particles were studied by stable-isotope labeling and high-throughput mass spectrometry. We used $^{16}O/^{18}O$ and iTRAQ labeling in

parallel approaches. $^{16}O/^{18}O$ labeling is expected to be more accurate in the linear ion trap mass spectrometer used, because the PQD fragmentation method in the low m/z region required to quantify iTRAQ reporters is less efficient; however, iTRAQ labeling has the advantage that four samples could be simultaneously compared, which was essential for the multiple comparisons needed to interpret the sources of variance. The quantitative data were interpreted using a statistical model developed previously in our laboratory [18], which decomposes the total quantification variance into three components: the spectrum variance (due to the error introduced when quantifying the isotope pairs in the mass spectrometer), the peptide variance (due to the error produced when peptides are prepared from their corresponding proteins) and the protein variance (due to the variability in protein composition and the errors introduced during the isolation of HDL and extraction of their proteins). The validity of this model for treating the data obtained by both stable isotope labeling methods is convincingly demonstrated by the accuracy with which the standardized variables corresponding to each variance component follow the expected null-hypothesis distributions in all the comparisons performed.

The decomposition of variance components was essential in order to identify the sources of variability encountered when analyzing the changes in HDL protein composition. Thus, by comparing variances at the spectrum and peptide levels with those obtained in other studies performed in our laboratory, we could convincingly demonstrate that the high variability observed when comparing HDL proteomes from different individuals was not due to technical problems or artifacts arising during peptide preparation and quantification, but instead, was due to HDL protein composition differences between different individuals. We were also able to conclude that the technical variability introduced by our HDL preparation procedure was very low, and therefore that the results were reproducible.

Using our statistical framework it was also possible to model a null-hypothesis distribution using a protein variance estimated from a different experiment. This allowed us to identify the proteins whose abundance differed significantly in HDL preparations from different individuals, taking into account the technical variability inherent to the HDL preparation procedure. Estimation of the null hypothesis in proteomics experiments is a non-trivial issue that deserves further consideration. In the vast majority of quantitative proteomics studies, the null hypothesis is estimated by assuming that the bulk of proteins is present at the same abundance in the two samples being compared. The distribution of quantitative data is then modeled on the basis of this assumption, so that a statistical significance can be assigned to each protein, producing a list of significant abundance changes. While this procedure is conservative, it has a very low sensitivity in cases where the most proteins in the two samples are differentially abundant, as is the case when HDL samples from different individuals are compared. The multiplexed nature of the iTRAQ approach made this method particularly appropriate for estimating the protein variance associated with the null hypothesis using our statistical model.

The statistical analysis allowed two main conclusions to be drawn. First, the HDL proteome shows very high inter-individual variability. This finding suggests that the HDL proteome composition is more dynamic than previously thought and reflects a fine tuning that is probably modulated predominantly by environmental factors. Second, the intra-individual protein composition of HDLs remains sufficiently stable after PTCA to reveal a consistent pattern of HDL protein abundance changes in different individuals, thus making it possible to determine a global pattern of protein alterations induced by PTCA.

Statistical analysis using threshold-free and systems biology approaches allowed the detection of a cluster of proteins implicated in HDL, cholesterol and phospholipid functions that were increased after PTCA. In addition to angiotensinogen and hepatic triacylglycerol lipase, the cluster was composed by apolipoproteins, including ApoC-II, ApoC-IV, and ApoA-IV. These three apolipoproteins are implicated in lipid transport and have a well-documented association with CAD [9,34]. ApoA-IV has also been implicated in the acute phase response and has anti-oxidative and anti-atherosclerotic properties. ApoA-IV is structurally and functionally similar to ApoA-I and has recently been shown to be redirected to HDL when triglyceride hydrolysis increases [35]. In other studies, ApoC-II and ApoA-IV, and possibly Apo C-IV, were shown to be required for maximal rates of triglyceride-rich lipoprotein lipolysis in HDL particles [36], and to modify different HDL functions, a finding consistent with our observation that PTCA increases hepatic triacylglycerol lipase. High concentrations of ApoC-II in HDL moreover impede reverse cholesterol transfer, thereby blocking HDL maturation [37]. A recent proteomic analysis found that HDL proteins related to lipid metabolism regulate sterol accumulation by macrophages, one of the most important cell types implicated in atherosclerosis development [38]. These results together indicate that the function of HDL particles can be finely regulated by their protein cargo, and the pattern of protein increases in HDL that we report here suggest an important interaction of HDL with atherosclerotic plaque, probably favoring lipolysis and fatty-acid release in injured tissues, hemostasis, and the interchange of substances between HDL and the content of ruptured plaques.

The other cluster of proteins that were increased by PTCA included proteins related to ATP and nucleotide binding, focal adhesion, and tight junctions, including several actin and myosin species. Among these, cytoplasmic actin1 (β-actin), is the main structural cytoskeletal constituent and is ubiquitously expressed in all eukaryotic cells. The additional cytoplasmic actin 1 accumulated in HDL after PTCA might come directly from atheroma plaque rupture, suggesting that HDL might carry waste cellular components from the plaque [9,12]. The liver-specific fibrinogen-like protein 1 was also significantly increased. Plasma fibrinogen is a biomarker of inflammation [39], and might be associated with the risk of coronary heart disease (CHD) and stroke [40]. In general, these increases are consistent with a putative role of HDL as a carrier of proteins from dead cells of the atheromatous plaque after PTCA surgery.

The proteins that were decreased by PTCA could not be clustered according to biological functions, but they gave further insight into the putative role of HDL in this process. The list of decreased proteins included antithrombin-III, also called serpin C1, which is the most important serine protease inhibitor in plasma, and regulates the blood coagulation cascade and presents anti-inflammatory properties [41].

Deficiency in antithrombin-III is associated with a high incidence of thrombotic events and a poor cardiovascular prognosis. HDL is known to be antithrombotic, due to its capacity to inhibit thrombin, platelet aggregation and the expression of the proinflammatory chemokine MCP1 [34,42]. Also reduced after PTCA were some immunoglobulins, including Igλ-1 chain C regions (IgLC1) and heavy chain variable region, and annexin A1. Annexin A1 is a phospholipase A2 inhibitor with potent anti-inflammatory activity [43]. IgLC1 is related to complement activation, and the immunoglobulin heavy chain variable region has been linked to significant abnormalities in HDL composition [44]. Based on these observations, we speculate that the protective properties of HDL might be impaired after PTCA, as a consequence of the decrease in the amount of antithrombin-III and of proteins related to innate immunity.

5. Conclusions

We have developed a technology that allows the systematic analysis of the HDL proteome and was applied to the study of HDL protein alterations resulting from the rupture of atheroma plaque. Our results are consistent with the multifunctional role of HDLs, and with the concept that HDL particles behave as complex platforms that donate and receive a variety of molecules. A well-known function of HDL is its ability to remove excess cholesterol from peripheral tissues and return it to the liver, a process called reverse cholesterol transport. The altered proportion of several proteins in HDL after PTCA strongly suggests that HDL particles are also involved in the transport of other substances. Furthermore, the inter-individual variability in protein content reported here probably reflects the complexity of this particle, which must adapt to individual stress situations. This would explain the high interaction of HDL with the immune system, probably as a consequence of its ability to function as a reservoir for immunoregulatory substances [45]. Given the high inter-individual variability and the sensitivity of HDL protein composition to angioplasty, our results suggest that our technology for HDL analysis could be used for the systematic study of differences among individuals and also as biomarkers for risk prediction of cardiovascular disease and other clinical conditions.

Conflict of interest

All authors declare no actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations that could inappropriately influence, or be perceived to influence their work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2014.04.010.

REFERENCES

[1] Assmann G, Cullen P, Schulte H. The Munster heart study (PROCAM). Results of follow-up at 8 years. Eur Heart J 1998;19(Suppl. A):A2–A11.

[2] Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. High density lipoprotein as a protective factor against coronary heart disease. The Framingham study. Am J Med 1977;62:707–14.

[3] Barter PJ, Caulfield M, Eriksson M, Grundy SM, Kastelein JJ, Komajda M. Effects of torcetrapib in patients at high risk for coronary events. N Engl J Med 2007;357:2109–22.

[4] Briel M, Ferreira-Gonzalez I, You JJ, Karanicolas PJ, Akl EA, Wu P, et al. Association between change in high density lipoprotein cholesterol and cardiovascular disease morbidity and mortality: systematic review and meta-regression analysis. BMJ 2009;338:b92.

[5] Burillo E, Andres EM, Mateo-Gallego R, Fiddyment S, Jarauta E, Cenarro A, et al. High-density lipoprotein cholesterol increase and non-cardiovascular mortality: a meta-analysis. Heart 2010;96:1345–51.

[6] Seo SM, Choo EH, Koh YS, Park MW, Shin DI, Choi YS, et al. Catholic University of Korea PrCIRi. High-density lipoprotein cholesterol as a predictor of clinical outcomes in patients achieving low-density lipoprotein cholesterol targets with statins after percutaneous coronary intervention. Heart 2011;97:1943–50.

[7] Brewer Jr HB. HDL metabolism and the role of HDL in the treatment of high-risk patients with cardiovascular disease. Curr Cardiol Rep 2007;9:486–92.

[8] Gordon SM, Deng J, Lu LJ, Davidson WS. Proteomic characterization of human plasma high density lipoprotein fractionated by gel filtration chromatography. J Proteome Res 2010;9:5239–49.

[9] Vaisar T, Pennathur S, Green PS, Gharib SA, Hoofnagle AN, Cheung MC, et al. Shotgun proteomics implicates protease inhibition and complement activation in the antiinflammatory properties of HDL. J Clin Invest 2007;117:746–56.

[10] Rau JC, Beaulieu LM, Huntington JA, Church FC. Serpins in thrombosis, hemostasis and fibrinolysis. J Thromb Haemost 2007;5(Suppl. 1):102–15.

[11] Heinecke JW. The protein cargo of HDL: implications for vascular wall biology and therapeutics. J Clin Lipidol 2010;4:371–5.

[12] Vaisar T. Proteomics investigations of HDL: challenges and promise. Curr Vasc Pharmacol 2012;10:410–21.

[13] Cheung MC, Segrest JP, Albers JJ, Cone JT, Brouillette CG, Chung BH, et al. Characterization of high density lipoprotein subspecies: structural studies by single vertical spin ultracentrifugation and immunoaffinity chromatography. J Lipid Res 1987;28:913–29.

[14] Bonzon-Kulichenko E, Perez-Hernandez D, Nunez E, Martinez-Acedo P, Navarro P, Trevisan-Herraz M, et al. A robust method for quantitative high-throughput analysis of proteomes by 18O labeling. Mol Cell Proteomics 2011;10 [M110 003335].

[15] Burillo E, Vazquez J, Jorge I. Quantitative proteomics analysis of high-density lipoproteins by stable (18)O-isotope labeling. Methods Mol Biol 2013;1000:139–56.

[16] Lopez-Ferrer D, Ramos-Fernandez A, Martinez-Bartolome S, Garcia-Ruiz P, Vazquez J. Quantitative proteomics using 16O/18O labeling and linear ion trap mass spectrometry. Proteomics 2006;6(Suppl. 1):S4–S11.

[17] Ramos-Fernandez A, Lopez-Ferrer D, Vazquez J. Improved method for differential expression proteomics using trypsin-catalyzed 18O labeling with a correction for labeling efficiency. Mol Cell Proteomics 2007;6:1274–86.

[18] Jorge I, Navarro P, Martinez-Acedo P, Nunez E, Serrano H, Alfranca A, et al. Statistical model to analyze quantitative proteomics data obtained by 18O/16O labeling and linear ion trap mass spectrometry: application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells. Mol Cell Proteomics 2009;8:1130–49.
[19] Lopez-Ferrer D, Martinez-Bartolome S, Villar M, Campillos M, Martin-Maroto F, Vazquez J. Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. Anal Chem 2004;76:6853–60.
[20] Martinez-Bartolome S, Navarro P, Martin-Maroto F, Lopez-Ferrer D, Ramos-Fernandez A, Villar M, et al. Properties of average score distributions of SEQUEST: the probability ratio method. Mol Cell Proteomics 2008;7:1135–45.
[21] Navarro P, Vazquez J. A refined method to calculate false discovery rates for peptide identification using decoy databases. J Proteome Res 2009;8:1792–6.
[22] Navarro P, Trevisan-Herraz M, Bonzon-Kulichenko E, Núñez E, Martínez-Acedo P, Pérez-Hernández D, et al. General statistical framework for quantitative proteomics by stable isotope labeling. J Proteome Res 2014;13:1234–47.
[23] Alwaili K, Bailey D, Awan Z, Bailey SD, Ruel I, Hafiane A, et al. The HDL proteome in acute coronary syndromes shifts to an inflammatory profile. Biochim Biophys Acta 1821;2012:405–15.
[24] Mange A, Goux A, Badiou S, Patrier L, Canaud B, Maudelonde T, et al. HDL proteome in hemodialysis patients: a quantitative nanoflow liquid chromatography–tandem mass spectrometry approach. PLoS One 2012;7:e34107.
[25] Heller M, Stalder D, Schlappritzi E, Hayn G, Matter U, Haeberli A. Mass spectrometry-based analytical tools for the molecular protein characterization of human plasma lipoproteins. Proteomics 2005;5:2619–30.
[26] Green PS, Vaisar T, Pennathur S, Kulstad JJ, Moore AB, Marcovina S, et al. Combined statin and niacin therapy remodels the high-density lipoprotein proteome. Circulation 2008;118:1259–67.
[27] Mazur MT, Cardasis HL, Spellman DS, Liaw A, Yates NA, Hendrickson RC. Quantitative analysis of intact apolipoproteins in human HDL by top-down differential mass spectrometry. Proc Natl Acad Sci U S A 2010;107:7728–33.
[28] Vaisar T, Mayer P, Nilsson E, Zhao XQ, Knopp R, Prazen BJ. HDL in humans with cardiovascular disease exhibits a proteomic signature. Clin Chim Acta 2010;411:972–9.
[29] Moore D, McNeal C, Macfarlane R. Isoforms of apolipoprotein C-I associated with individuals with coronary artery disease. Biochem Biophys Res Commun 2011;404:1034–8.
[30] Davidson WS, Silva RA, Chantepie S, Lagor WR, Chapman MJ, Kontush A. Proteomic analysis of defined HDL subpopulations reveals particle-specific protein clusters: relevance to antioxidative function. Arterioscler Thromb Vasc Biol 2009;29:870–6.
[31] Rezaee F, Casetta B, Levels JH, Speijer D, Meijers JC. Proteomic analysis of high-density lipoprotein. Proteomics 2006;6:721–30.
[32] Havel RJ, Eder HA, Bragdon JH. The distribution and chemical composition of ultracentrifugally separated lipoproteins in human serum. J Clin Invest 1955;34:1345–53.
[33] Asztalos BF, Tani M, Schaefer EJ. Metabolic and functional relevance of hdl subspecies. Curr Opin Lipidol 2011;22:176–85.
[34] Tölle M, Huang T, Schuchardt M, Jankowski V, Prüfer N, Jankowski J, et al. High-density lipoprotein loses its anti-inflammatory capacity by accumulation of pro-inflammatory-serum amyloid A. Cardiovasc Res 2012;94:154–62.
[35] Duka A, Fotakis P, Georgiadou D, Kateifides A, Tzavlaki K, von Eckardstein L, et al. ApoA-IV promotes the biogenesis of ApoA-IV-containing HDL particles with the participation of ABCA1 and LCAT. J Lipid Res 2013;54:107–15.
[36] Goldberg IJ, Scheraldi CA, Yacoub LK, Saxena U, Bisgaier CL. Lipoprotein ApoC-II activation of lipoprotein lipase. Modulation by apolipoprotein A-IV. J Biol Chem 1990;265:4266–72.
[37] Tian L, Xu Y, Fu M, Jia L, Yang Y. Influence of apolipoprotein CII concentrations on hdl subclass distribution. J Atheroscler Thromb 2009;16:611–20.
[38] Suzuki M, Becker L, Pritchard DK, Gharib SA, Wijsman EM, Bammler TK, et al. Cholesterol accumulation regulates expression of macrophage proteins implicated in proteolysis and complement activation. Arterioscler Thromb Vasc Biol 2012;32:2910–8.
[39] Ross R. Atherosclerosis—an inflammatory disease. N Engl J Med 1999;340:115–26.
[40] Farrell DH. Γ′ fibrinogen as a novel marker of thrombotic disease. Clin Chem Lab Med 2012;50:1903–9.
[41] Afshari A, Wetterslev J, Brok J, Moller A. Antithrombin III in critically ill patients: systematic review with meta-analysis and trial sequential analysis. BMJ 2007;335:1248–51.
[42] Nofer JR, Kehrel B, Fobker M, Levkau B, Assmann G, von Eckardstein A. HDL and arteriosclerosis: beyond reverse cholesterol transport. Atherosclerosis 2002;161:1–16.
[43] Viiri LE, Full LE, Navin TJ, Begum S, Didangelos A, Astola N, et al. Smooth muscle cells in human atherosclerosis: proteomic profiling reveals differences in expression of annexin A1 and mitochondrial proteins in carotid disease. J Mol Cell Cardiol 2013;54:65–72.
[44] Mendez AJ, Goldberg RB, Arnold PI, Schultz DR. Acquire HDL deficiency associated with apolipoprotein A-I reactive monoclonal immunoglobulins. Arterioscler Thromb Vasc Biol 2002;22.
[45] Norata GD, Pirillo A, Ammirati E, Catapano AL. Emerging role of high density lipoproteins as a player in the immune system. Atherosclerosis 2012;220:11–21.

1.2 Quantitative HDL proteomics identifies peroxiredoxin-6 as a biomarker of human abdominal aortic aneurysm

(Information on this publication is in subsection 4.3 of the Results chapter.)

SCIENTIFIC REPORTS

OPEN

Quantitative HDL Proteomics Identifies Peroxiredoxin-6 as a Biomarker of Human Abdominal Aortic Aneurysm

Received: 08 August 2016
Accepted: 09 November 2016
Published: 09 December 2016

Elena Burillo[1,*], Inmaculada Jorge[2,*], Diego Martínez-López[1], Emilio Camafeita[2], Luis Miguel Blanco-Colio[1], Marco Trevisan-Herraz[2], Iakes Ezkurdia[2], Jesús Egido[3], Jean-Baptiste Michel[4], Olivier Meilhac[5], Jesús Vázquez[2,*] & Jose Luis Martin-Ventura[1,*]

High-density lipoproteins (HDLs) are complex protein and lipid assemblies whose composition is known to change in diverse pathological situations. Analysis of the HDL proteome can thus provide insight into the main mechanisms underlying abdominal aortic aneurysm (AAA) and potentially detect novel systemic biomarkers. We performed a multiplexed quantitative proteomics analysis of HDLs isolated from plasma of AAA patients (N = 14) and control study participants (N = 7). Validation was performed by western-blot (HDL), immunohistochemistry (tissue), and ELISA (plasma). HDL from AAA patients showed elevated expression of peroxiredoxin-6 (PRDX6), HLA class I histocompatibility antigen (HLA-I), retinol-binding protein 4, and paraoxonase/arylesterase 1 (PON1), whereas α-2 macroglobulin and C4b-binding protein were decreased. The main pathways associated with HDL alterations in AAA were oxidative stress and immune-inflammatory responses. In AAA tissue, PRDX6 colocalized with neutrophils, vascular smooth muscle cells, and lipid oxidation. Moreover, plasma PRDX6 was higher in AAA (N = 47) than in controls (N = 27), reflecting increased systemic oxidative stress. Finally, a positive correlation was recorded between PRDX6 and AAA diameter. The analysis of the HDL proteome demonstrates that redox imbalance is a major mechanism in AAA, identifying the antioxidant PRDX6 as a novel systemic biomarker of AAA.

Abdominal aortic aneurysm (AAA) is a major health problem, with a prevalence of ~2% in adults aged over 65 years[1,2] and causing about 1–2% of male deaths in economically developed societies[3]. Clinically, AAA is defined as a permanent dilation of the aortic diameter by more than 3 cm or more than 50% of the initial value. Mechanistically, AAA is characterized by the formation of an intraluminal thrombus (ILT), proteolysis, oxidative stress, immune inflammatory response, angiogenesis and fibrosis. Currently, the only way to prevent aortic rupture in patients with an AAA >5.5 cm is surgery. AAA is usually asymptomatic and is often detected as an incidental finding during the investigation of an unrelated problem or as a consequence of radiological screening. Moreover, diameter growth is discontinuous, with periods of growth alternating with periods of stability, making prognosis difficult[4]. There is thus a pressing need to identify novel biomarkers of the presence and evolution of AAA, which provide insight into the pathological mechanisms of the disease.

The most convenient source of a systemic AAA biomarker is blood plasma. Our group and others have previously analyzed AAA patient plasma using a variety of proteomics approaches[5–9]. However, the high dynamic range of protein concentrations in plasma makes it difficult to quantify proteins present in low amounts, even after depletion of the most abundant proteins. To circumvent this problem, some authors have concentrated on the analysis of specific plasma subproteomes, such as high-density lipoproteins (HDLs).

[1]Vascular Research Lab, IIS-Fundación Jiménez Díaz-Autonoma University, Madrid, Spain. [2]Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain. [3]Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Spain. [4]Inserm, U698, Universite Paris 7, CHU X-Bichat, Paris, France. [5]Diabète athérothrombose Thérapies Réunion Océan Indien (UMR DéTROI U1188) – Université de La Réunion-CYROI- 2, rue Maxime Rivière 97490 Sainte Clotilde – La Réunion – France. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.V. (email: jvazquez@cnic.es) or J.L.M.-V. (email: jlmartin@fjd.es)

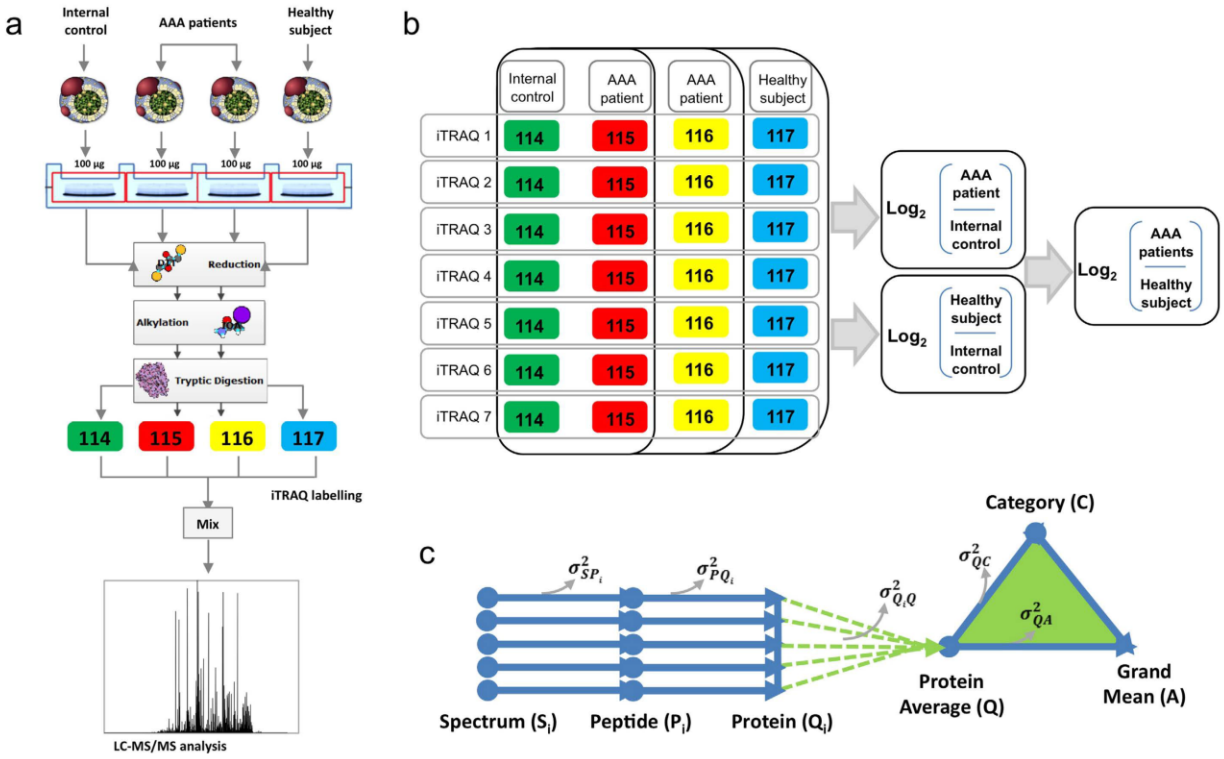


Figure 1. Multiplexed analysis of individual HDL proteomes from AAA patients: shotgun proteomics workflow. (**a**) HDL particles were isolated by sequential ultracentrifugation. Protein extracts were loaded onto SDS-PAGE gels, and proteins were concentrated in the stacking gel. After in-gel trypsin digestion, peptides were labeled with 4-plex iTRAQ reagents and then analyzed by LC-MS/MS. (**b**) A total of 7 iTRAQ experiments were performed. In each experiment, samples from 2 AAA patients and 1 control subject were compared with an internal control. The internal control was a pool of all the individual samples used in the study. (**c**) The statistical model for the quantitative data decomposes the total technical variance into the spectral, peptide, and protein variance components and provides a general framework to fully integrate quantitative and error information, allowing a comparative analysis of the results obtained from the 7 iTRAQ experiments.

HDLs are a transport platform of lipids and proteins. HDLs are responsible for transporting excess cholesterol from peripheral tissues to the liver for its elimination in feces and bile. Clinical and epidemiological studies consistently link low HDL cholesterol levels to an elevated risk of cardiovascular disease. A recent meta-analysis also revealed a negative association of HDL cholesterol levels with AAA in most studies[10]. HDLs are additionally a transport platform for constitutive and non-permanently associated plasma proteins. Screening of the HDL proteome has identified dramatic protein alterations in a variety of disease contexts, including cardiovascular diseases[11–15]. HDLs have several important cardiovascular protective properties, including anti-oxidant, anti-inflammatory, and antithrombotic effects[16,17]. These properties are attributable to HDL-associated proteins, and it is increasingly accepted that HDL composition, rather than quantity, is more important for its vasculo-protective activities. We previously demonstrated impaired anti-proteolytic[18] and anti-oxidative functions[19] in HDLs from human AAA. This effect has been linked to altered HDL protein composition or to impaired function of individual components. However, the protein alterations taking place in the HDL proteome in AAA have not been reported to date.

Here, we performed a multiplexed quantitative proteomics study of HDL-protein alterations in AAA. Systems biology analysis revealed an association of AAA with an increased HDL content of proteins related to redox homeostasis, most clearly evident in the increased levels of peroxiredoxin-6 (PRDX6). Circulating levels of PRDX6 are also increased in AAA-patient plasma, and PRDX6 levels correlate positively with AAA size, identifying PRDX6 as a promising biomarker of AAA.

Results

Analysis of the human HDL proteome in AAA. HDL particles were isolated from plasma and subjected to trypsin digestion. The resulting peptides were iTRAQ-labeled and combined in 7 independent experiments for LC-MS/MS analysis (Fig. 1). The protein composition of the HDL samples was characterized by spectral counting of the 7 HDL pools. At a 1% FDR threshold, 112 proteins were identified as HDL particle constituents in at least 3 experiments (Supplementary Table S1). As expected, ApoA1 was one of the most abundant HDL proteins, representing 33% of total protein composition. Other abundant HDL proteins were α-1-antitrypsin, complement C3, apolipoproteins D, E and M, and paraoxonase/arylesterase 1 (3–4% each) (Supplementary Figure S1). Albumin and ApoB100 were also identified but were not included in the analysis because they are abundant

	Gene name	Accession number	Protein name	Xq = Log2 (AAA/control) Individual comparatives 1	2	3	4	5	6	7	8	9	10	11	12	13	14	Integration data Xq	Zq	p-value	FDR
Redox/homeostasis- related proteins	PRDX6	**P30041**	**PRDX6_HUMAN Peroxiredoxin-6**	1.12		0.77	1.78				1.59		1.82	0.58				**1.29**	**3.79**	**0.0002**	**0.0116**
	RBP4	**P02753**	**RET4_HUMAN Retinol-binding protein 4**	1.71	2.32	1.02	1.45	1.43	1.89	2.26	0.35	0.71	0.75	0.69	0.82	-0.77	-0.29	**1.03**	**3.59**	**0.0003**	**0.0176**
	PON1	**P27169**	**PON1_HUMAN Serum paraoxonase/arylesterase 1**	0.16	1.24	1.88	0.95	1.02	1.22	1.55	0.57	0.47	0.03	1.74	1.58	-0.05	1.01	**0.95**	**3.35**	**0.0008**	**0.0309**
	CST3	P01034	CYTC_HUMAN Cystatin-C	0.42	1.56	1.09	1.19	1.38	2.82	2.66	0.69	0.66	1.01	0.54	1.11	1.69	1.37	0.87	**2.98**	**0.0029**	0.0847
	APOA4	P06727	APOA4_HUMAN Apolipoprotein A-IV	0.50	1.11	0.73	0.59	1.00	0.74	0.21	-0.13	-0.19	0.18	-0.25	-0.04	-0.13	-0.60	0.26	1.04	0.2976	0.9804
	HBA1	P69905	HBA_HUMAN Hemoglobin subunit alpha	-0.41	-0.45	0.54	-0.02	0.29	0.37	0.25	0.62	0.02	0.05	0.09	0.20	2.55	-0.06	0.26	1.00	0.3196	0.9967
	HBB	P68871	HBB_HUMAN Hemoglobin subunit beta	-0.02	0.08	2.12	0.12	0.18	-0.25	0.00	1.00	-0.03	0.01	0.53	-0.49	3.29	0.93	0.11	0.50	0.6155	1.0074
	PON3	Q15166	PON3_HUMAN Serum paraoxonase/lactonase 3	0.04	0.33	0.18	-0.07	-0.54	-0.25	-0.15	0.23	0.42	0.61	-0.09	-0.01	0.27	-0.06	0.05	0.33	0.7432	1.0017
	HP	P00738	HPT_HUMAN Haptoglobin	-0.47	-1.65	0.31	1.04	2.35	0.86	-0.15	-0.84	-1.18	0.67	0.89	0.47	1.58	1.26	-0.03	0.05	0.9629	1.0009
MHC class I protein complex	HLA-A23	**P30447**	**1A23_HUMAN HLA class I histocompatibility antigen, A-23 alpha chain**	-1.98		-0.53		1.26	1.83		-1.65		1.34		3.61	2.74		**1.13**	**3.51**	**0.0004**	**0.0211**
	B2M	P61769	B2MG_HUMAN Beta-2-microglobulin	0.41	1.55	0.63	0.29	0.65	0.36	0.16	0.24	0.63	1.04	0.06	0.46	1.77	0.01	0.61	2.12	**0.0339**	0.4253
	HLA-A1	P30443	1A01_HUMAN HLA class I histocompatibility antigen, A-1 alpha chain	-1.77	-0.33	0.06	1.21	1.04	1.20	0.02	-0.88	0.08	1.64	0.39	2.37	1.82	-0.19	0.33	1.21	0.2263	0.9391
	UBA52	P62987	RL40_HUMAN Ubiquitin-60S ribosomal protein L40		0.37	0.20	0.16	-0.21	1.35	0.92		0.32	0.26	0.04	0.90	-0.47	0.03	0.33	1.18	0.2381	0.9578
	AZGP1	P25311	ZA2G_HUMAN Zinc-alpha-2-glycoprotein		0.93		0.67	0.84	0.75	0.22		0.03		-0.44	1.59	0.39	-0.72	0.16	0.62	0.5376	0.9670
Acute phase response	SAA	E9PQD6	E9PQD6_HUMAN Serum amyloid A protein	-0.31	1.82	-0.65				1.52	0.00	-0.28						0.94	2.78	**0.0055**	0.1454
	LBP	P18428	LBP_HUMAN Lipopolysaccharide-binding protein	0.36	0.14	0.68	1.25	1.46	0.57	0.36	0.17	0.72	-0.11	0.42	1.46	0.28	0.32	0.55	1.94	0.0530	0.5582
	SAA	P0DJI8	SAA1_HUMAN Serum amyloid A-1 protein	1.86	1.60	-1.97	0.23	-0.22	-0.71	-0.44	0.45	0.33	-0.58	0.17	0.40	2.24	-0.11	0.44	1.59	0.1124	0.8227
	AHSG	P02765	FETUA_HUMAN Alpha-2-HS-glycoprotein	0.51	0.95	1.15	0.85	1.20	-0.25	1.37	-0.47	-0.06	-0.17	-0.07	-0.05	0.07	-0.04	0.35	1.33	0.1838	0.9139
	SAA	D3DQX7	D3DQX7_HUMAN Serum amyloid A protein	0.36		-0.86	-0.24	0.08	-0.23	-0.66	1.15		-0.23	0.63	0.70	-0.31	0.47	0.29	1.05	0.2946	0.9826
	SAA2	P0DJI9	SAA2_HUMAN Serum amyloid A-2 protein	1.44	0.55	-1.10	0.64		-1.10		0.34	-0.02	-0.59	-0.14		0.86		0.26	0.95	0.3408	1.0146
	ITIH4	Q14624	ITIH4_HUMAN Inter-alpha-trypsin inhibitor heavy chain H4	-1.14	-1.71	-0.27	2.08	1.07	2.52		-0.38	0.63	-1.53	-0.19	-0.65	2.55		0.12	0.52	0.6012	1.0122
	ORM1	P02763	A1AG1_HUMAN Alpha-1-acid glycoprotein 1	-0.71	0.42	-0.52	-1.42	1.14	1.69	0.89	-0.35	-0.07	-0.22	-1.77	0.44	-0.67	-0.09	0.12	0.52	0.6015	1.0095
	ORM2	P19652	A1AG2_HUMAN Alpha-1-acid glycoprotein 2	-0.49	-0.15	0.03	0.35	0.73	1.63	1.42	-0.22	-0.31	-0.22	0.32	-0.07	-0.32	-0.35	0.09	0.42	0.6745	1.0098
	SAA	B2R5G8	B2R5G8_HUMAN Serum amyloid A protein	0.71	0.79	-0.21	-0.20	-0.82	-0.39	-0.67	0.62	-0.21	0.11	0.13	-0.46	1.06	0.01	0.03	0.24	0.8128	1.0056
	SAA4	P35542	SAA4_HUMAN Serum amyloid A-4 protein	0.42	0.48	-0.21	-0.32	-0.68	-0.54	-0.06	0.42	-0.85	0.24	0.09	-0.07	1.21	0.02	0.00	0.15	0.8771	1.0137
	HP	P00738	HPT_HUMAN Haptoglobin	-0.47	-1.65	0.31	1.04	2.35	0.86	-0.15	-0.84	-1.18	0.67	0.89	0.47	1.58	1.26	-0.03	0.05	0.9629	1.0009
	SERPINA1	P01009	A1AT_HUMAN Alpha-1-antitrypsin	-0.09	0.74	0.37	0.06	0.97	0.22		-1.51	-0.44	-0.22	-1.21	-0.38	-1.25		-0.09	-0.16	0.8717	1.0141
	SERPINF2	P08697	A2AP_HUMAN Alpha-2-antiplasmin	-0.04		-0.08		-0.55			-0.04		-0.13		-0.65			-0.26	-0.62	0.5338	0.9701
Platelet-related proteins	PFN1	P07737	PROF1_HUMAN Profilin-1	1.16		0.11	1.30	2.04	1.24	-0.39	1.10		1.57	0.48	1.12	0.14	1.10	0.90	3.04	**0.0024**	0.0730
	PF4	P02776	PLF4_HUMAN Platelet factor 4			0.06	0.65	0.51	1.84				1.90	0.90	-0.23	0.40		0.79	2.55	**0.0108**	0.2370
	PPBP	P02775	CXCL7_HUMAN Platelet basic protein	0.60	0.43	0.12	0.44	0.84	1.44	-0.07	1.55	0.04	1.20	0.38	-0.19	1.14	0.37	0.60	2.13	**0.0331**	0.4364
	RAP1A	P62834	RAP1A_HUMAN Ras-related protein Rap-1A		-0.09	0.00	2.10	1.24	0.52			0.86	3.33	1.52	1.00	-2.12		0.65	1.98	**0.0479**	0.5151
	PLA2G7	Q13093	PAFA_HUMAN Platelet-activating factor acetylhydrolase	1.63	0.95	0.97	-0.47	-0.86	-0.08	-0.62	0.23	0.37	0.74	0.47	0.48	1.24	1.07	0.43	1.54	0.1244	0.8192
	PF4V1	P10720	PF4V_HUMAN Platelet factor 4 variant	-0.25	0.35	-0.09	0.10	0.38	1.62		0.47	0.33	0.60	0.18	0.46	-0.07		0.42	1.47	0.1411	0.8171
Other increased proteins	ACTA2	P62736	ACTA_HUMAN Actin, aortic smooth muscle	0.31	0.16	-0.06	1.10	1.10	2.23	-1.03	1.25	1.31	1.87	0.82	1.47	2.28	0.93	0.77	2.69	**0.0072**	0.1724
	CNDP1	Q96KN2	CNDP1_HUMAN Beta-Ala-His dipeptidase	1.41	0.99	-1.65	1.41	1.41	1.27	1.62	0.16	0.54	-0.36	0.62	-0.02	-0.53	0.48	0.72	2.52	**0.0117**	0.2474
	LPA	Q1HP67	Q1HP67_HUMAN Lipoprotein, Lp(A)	2.36	3.39		-2.12		0.10		-0.60	0.16				-0.17		0.79	2.44	**0.0148**	0.2893
	SERPINF1	P36955	PEDF_HUMAN Pigment epithelium-derived factor	0.83	1.32	0.66	1.59	1.92	1.78	1.34	-0.30	0.18	0.04	0.01	0.45	-0.60	-0.16	0.64	2.28	**0.0226**	0.3507
	GC	P02774	VTDB_HUMAN Vitamin D-binding protein	0.58	1.08	0.61	1.15	2.17	1.52	1.96	-0.16	-0.21	0.11	-0.15	0.34	0.00	-0.10	0.64	2.27	**0.0235**	0.3539
	LOC613037	A8MRT5	K220L_HUMAN Putative NPIP-like protein LOC100132247		1.50	0.73		0.66	0.15			1.47	0.44		0.92	-1.78		0.69	2.18	**0.0294**	0.4082
	LCAT	P04180	LCAT_HUMAN Phphatidylcholine-sterol acyltransferase	0.51	1.05	0.56	0.73	1.15	1.58	2.23	0.13	0.46	0.02	0.46	0.96	-0.45	0.56	0.58	2.08	**0.0372**	0.4458
	AMBP	P02760	AMBP_HUMAN Protein AMBP	0.89	1.37	0.87	0.47	1.34	1.31	1.23	-0.14	0.12	0.58	-0.25	0.31	0.41	-0.36	0.57	2.04	**0.0411**	0.4608
	N/A	B3KNA1	B3KNA1_HUMAN cDNA FLJ14021 fis, clone HEMBA1002513, highly similar to HDAC6	0.16	-0.25				0.24	0.61	-0.28					-0.55		0.62	1.87	0.0611	0.5853
	ACTB	P60709	ACTB_HUMAN Actin, cytoplasmic 1	-0.08	0.24	-0.17	0.88	0.53	0.43	-0.67	0.88	0.79	-1.22	0.69	0.62	0.85	0.75	0.47	1.68	0.0928	0.7193
	PGBD3	Q8N328	PGBD3_HUMAN PiggyBac transposable element-derived protein 3		0.66	1.22			1.22			0.27	0.02			0.10		0.59	1.78	0.0755	0.6522
Other decreased proteins	DSG1	Q02413	DSG1_HUMAN Desmoglein-1		1.42	-0.01	-0.32			-1.11		0.22	-0.29	0.05			-0.64	-0.54	-1.53	0.1259	0.8193
	APOH	P02749	APOH_HUMAN Beta-2-glycoprotein 1	-0.44	-0.72	0.06	-0.53	-0.78	-0.84	-0.38	0.09	0.00	1.29	-0.43	-0.71	-0.69	-1.26	-0.56	-1.75	0.0805	0.6632
	IGLL5	B9A064	IGLL5_HUMAN Immunoglobulin lambda-like polypeptide 5	-0.12	-1.42		-0.24	-0.61	-1.10	-1.49	-0.61	-0.34		0.25	-0.98	-1.04	-0.45	-0.62	-1.90	0.0572	0.5685
	MMRN2	Q9H8L6	MMRN2_HUMAN Multimerin-2 OS					-1.15	-1.05	-0.81					-0.54	-0.15	-0.34	-0.72	-1.91	0.0562	0.5698
	C2orf15	Q8WU43	CB015_HUMAN Uncharacterized protein C2orf15	-0.01					-1.90	-2.74	-0.02					-0.17	-0.71	-0.85	-2.23	**0.0259**	0.3696
	IGHM	P01871	IGHM_HUMAN Ig mu chain C region	-0.18	-0.62	-0.49	-0.13	-2.31	0.20	-1.66	-0.38	-1.18	-0.04	-0.23	-2.52	-0.06	-1.02	-0.71	-2.23	**0.0256**	0.3752
	TF	P02787	TRFE_HUMAN Serotransferrin	-0.32	-1.04	-0.07		-1.04	-2.17	-1.84	-0.45	-1.01	-0.36		-0.99	-0.64	-0.50	-0.84	-2.86	**0.0079**	0.1800
	A2M	**P01023**	**A2MG_HUMAN Alpha-2-macroglobulin**	-0.89		0.03		-3.22	-1.86		-1.11		-0.26		-2.77	-1.33		-1.49	-4.47	**0.0000**	**0.0008**
Complement activation, classical pathway	C9	P02748	CO9_HUMAN Complement component C9	0.66	-0.09	-0.66	0.69	0.67	-0.16	-0.32	0.34	0.24	0.00	0.42	0.33	0.45	0.49	0.20	0.84	0.4024	1.0195
	C4A	P0C0L4	CO4A_HUMAN Complement C4-A	-0.85	-0.02	0.41	0.32	0.18	1.60	0.58	0.16	0.12	0.06	-0.13	-0.11	-0.19	-0.58	0.03	0.25	0.8046	1.0072
	SERPING1	P05155	IC1_HUMAN Plasma protease C1 inhibitor	0.35	-0.91	-1.30	-0.36	-1.15	1.25	-0.42	1.15	0.40	-0.51	-0.86	-0.51	-0.06	-1.06	-0.29	-0.84	0.3983	1.0140
	CLU	P10909	CLUS_HUMAN Clusterin	-0.22	-0.39	0.00	-0.17	-0.27	-0.56	-0.30	-0.04	-0.08	-0.36	-0.44	-0.36	-0.33	-0.56	-0.30	-0.90	0.3666	1.0223
	IGKC	P01834	IGKC_HUMAN Ig kappa chain C region	0.06	-0.82	0.15	-0.38	-0.40	-0.94	-0.66	-0.19	-0.57	0.59	-0.31	-0.45	-0.06	-0.35	-0.31	-0.92	0.3556	1.0186
	IGHG2	P01859	IGHG2_HUMAN Ig gamma-2 chain C region	-0.18		-0.59	-0.56	-1.23	-1.67	-1.31	-0.50		-0.13	0.39	-1.02	0.36	-0.04	-0.52	-1.53	0.1263	0.8018
	IGHG1	P01857	IGHG1_HUMAN Ig gamma-1 chain C region	-0.49	-1.46	0.17	-0.55	-0.60	-1.55	-1.71	-0.39	-0.26	1.44	-0.51	-0.16	-0.38	-0.35	-0.49	-1.56	0.1196	0.8078
	C3	P01024	CO3_HUMAN Complement C3	-0.80	-0.38	-0.03	-0.61	-0.35	-0.71	-0.88	-0.63	-0.54	-0.68	-0.54	-0.77	-0.72	-1.44	-0.62	-2.04	**0.0412**	0.4523
	C4BPA	**P04003**	**C4BPA_HUMAN C4b-binding protein alpha chain**	-1.50	-2.03		-1.54	-2.06		-1.21	-0.95	-1.22		-0.94	-1.36		-0.49	-1.36	-4.26	**0.0000**	**0.0018**

decrease in AAA: -3 -2 -1 0 1 2 3 increase in AAA

Figure 2. Protein abundance changes in the HDL proteome of AAA patients. Quantitative data for proteins belonging to significantly altered categories according to the colour scale at the bottom. Statistically significant changes ($p\text{-values} < 0.05$; $FDR < 0.05$) are shown in bold.

plasma proteins and are considered concomitant contaminants of HDLs obtained by ultracentrifugation[12]. Of the proteins identified in this study, 78 have previously been described as HDL-associated (Supplementary Table S1). These proteins comprise 92% of the absolute HDL protein content. A further 34 HDL proteins identified here, representing 8% of total HDL composition, have not been previously linked to HDL (Supplementary Table S1).

The complete list of quantitative results at the protein level is presented in Supplementary Table S2. Within the set of proteins found in at least 3 iTRAQ experiments, α-2 macroglobulin and C4b-binding protein were the two proteins showing the most prominent abundance deficit in HDL from AAA patients compared with controls, whereas PRDX6, HLA-I, retinol-binding protein 4, and PON1 were the proteins showing the largest abundance increases (Fig. 2).

Among the identified proteins, PRDX6 attracted our interest because it showed the most significant difference between AAA patients and controls and had not previously been linked to HDL or AAA. The quantitative proteomics results for PRDX6 were therefore validated by western blot in the same 14 AAA and 7 control samples (Fig. 3a and Supplementary Figure S2). Western blot was also used to validate HLA-I and PON1, which were also not previously linked to AAA (Fig. 3b and Supplementary Figure S2).

To explore whether the observed protein changes were discrete events or part of a more general pattern of HDL protein alterations, we performed a systems biology analysis using the SBT model, a statistical framework that captures functional alterations produced by coordinated protein abundance changes[20]. SBT analysis was performed on the averaged protein quantifications obtained from the integration of the experiments. We detected significant increases in proteins related to redox homeostasis ($p = 0.00001$, $FDR = 0.00045$), MHC class I complex ($p = 0.00178$, $FDR = 0.01386$), acute phase response ($p = 0.03597$, $FDR = 0.06570$), and platelets ($p = 0.00002$, $FDR = 0.00076$), whereas complement activation proteins from the classical pathway were decreased ($p = 0.00042$, $FDR = 0.00787$) (Fig. 2). With the exception of α-2 macroglobulin, all the proteins detected individually as being significantly altered were found among these altered functional categories, suggesting that they were the most evident consequence of a general reconfiguration of the HDL proteome. This interpretation was reinforced by

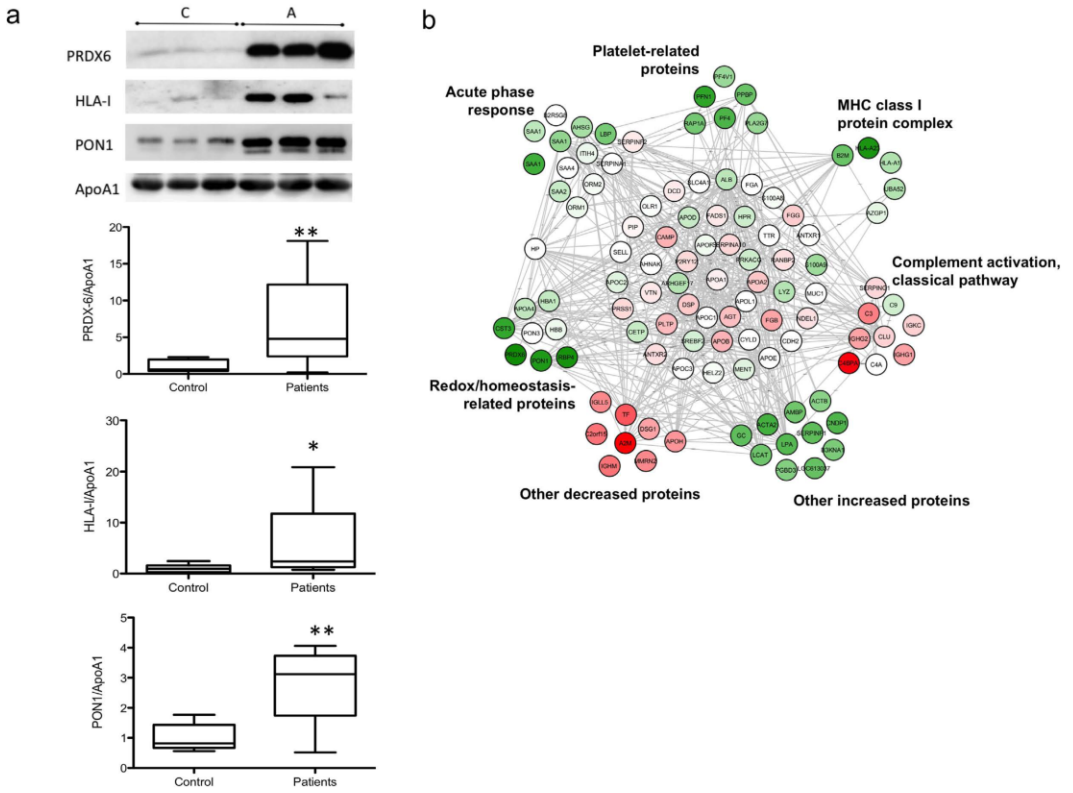


Figure 3. Validation and interaction network of the AAA HDL proteome. (**a**) Representative protein validation by western blot. The quantitative results for PRDX6, HLA-I, and PON1 correspond to the same 14 AAA and 7 control samples used in the proteomic analysis. *$p < 0.05$, **$p < 0.01$. (**b**) Interaction network of HDL proteins showing the quantitative data. Significant category changes were clustered.

network analysis, which indicated that most of the altered HDL proteins were interconnected in a dense interaction network containing most of the quantified HDL proteins (Fig. 3b).

Tissue expression of PRDX6 in AAA. Of the validated proteins, we focused on PRDX6 because it reflects the generalized alteration in proteins implicated in redox homeostasis, the functional category most clearly changed in the HDL proteome of AAA patients. Immunohistochemical analysis of PRDX6 expression and localization in the ILT obtained from AAA patients revealed intense PRDX6 immunostaining associated with areas of neutrophil infiltration and RBC degeneration (Figs 4 and 5). Examination of the aortic wall of AAA patients showed strong PRDX6 staining in cholesterol-rich and acellular atherosclerotic plaques in the media layer, but also revealed PRDX6 colocalization with vascular smooth muscle cells (VSMCs) (Fig. 4). PRDX6 also colocalized in AAA tissue with the lipid peroxidation marker MDA and with ceroids, markers of RBC-associated oxidation (Fig. 5).

Circulating PRDX6 in AAA patients. The association of PRDX6 with circulating HDL particles could arise from interaction of HDL particles with circulating cells or from direct loading from plasma. The localization of PRDX6 in areas of erythrophagocytosis suggested that the presence of PRDX6 in HDL particles could arise, at least in part, from the interaction of HDLs with circulating RBCs. To test this hypothesis, we incubated HDLs isolated from healthy subjects with intact or lysed RBCs, and then re-isolated the HDLs to analyse PRDX6 content. PRDX6 levels were low in untreated HDLs from healthy subjects, but increased markedly when incubated with intact or lysed RBCs (Fig. 6). In control experiments, the abundant RBC protein catalase was taken up by HDL particles incubated with lysed RBCs but not by particles incubated with intact RBCs.

Given that HDL is a platform of plasma-associated proteins, we also examined whether PRDX6 is present in plasma and could serve as a circulating biomarker of AAA. The plasma concentration of PRDX6 was analyzed in a cohort of control individuals ($N = 27$) and AAA patients ($N = 47$). Cohort clinical characteristics are shown in Table 1. This analysis confirmed the presence of PRDX6 in plasma, which to our knowledge has not been reported before, and also revealed higher plasma PRDX6 levels in AAA patients than in controls (21 ± 2 vs 10 ± 2 ng/mL; $p < 0.01$) (Fig. 7a). Multivariate logistic regression analysis of AAA risk factors revealed plasma PRDX6 concentration, hypertension, and smoking to be independent predictors of the presence of AAA [Odds Ratio CI (95%): 1.060 (1.006–1.117), $p < 0.05$ for PRDX6; 5.775 (1.670–19.973), $p < 0.01$ for hypertension; and 10.028 (1.139–88.311), $p < 0.05$ for smoking)]. Furthermore, plasma PRDX6 level correlated positively with AAA size ($r = 0.4$, $p < 0.001$, adjusted for age) (Fig. 7b), reinforcing the potential of PRDX6 as a biomarker of AAA.

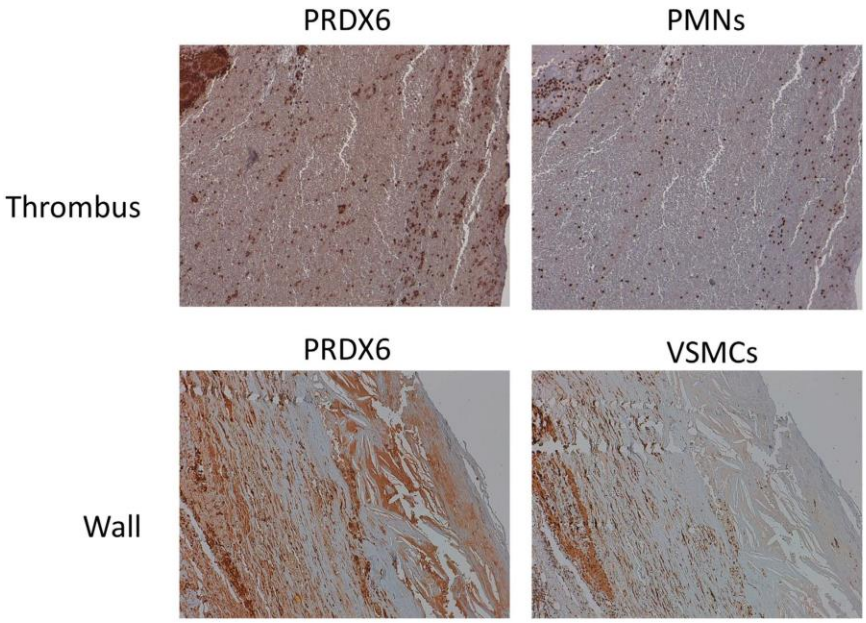


Figure 4. PRDX6 immunohistochemistry in AAA tissue. PRDX6 was detected in AAA thrombus (colocalizing with neutrophils: PMNs, CD15 staining) and wall (colocalizing with vascular smooth muscle cells: VSMC, alpha actin staining). Magnification x10. N = 10.

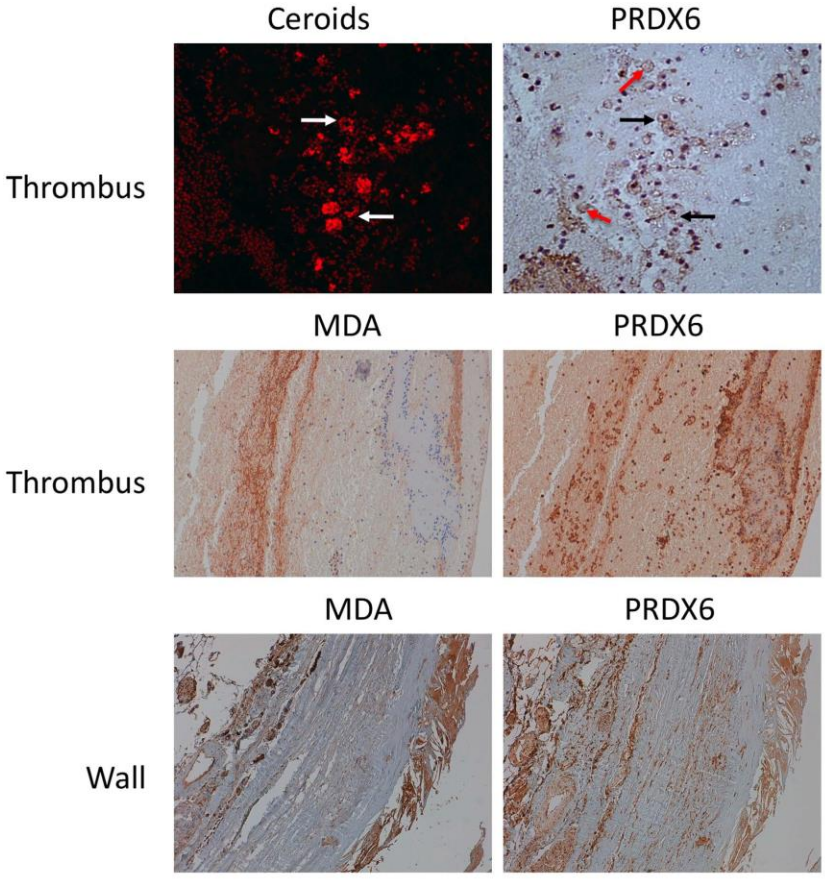


Figure 5. PRDX6 is expressed in areas of high oxidative stress in AAA tissue. PRDX6 colocalizes with ceroids and the lipid peroxidation marker MDA in AAA thrombus and wall. Black arrow indicates typical ceroid rings; red arrow indicates degenerated red blood cells. Magnification x10 (x40 in the upper part of ceroids and PRDX6). N = 10.

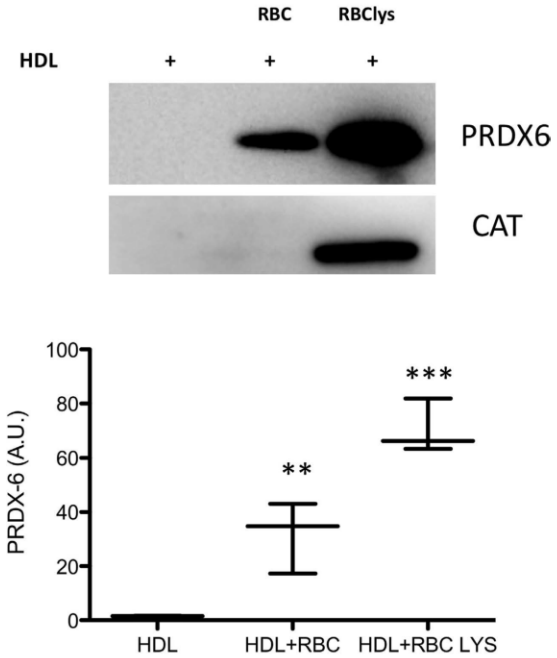


Figure 6. PRDX6 protein levels in HDL after incubation with RBCs. Western blot of PRDX6 and catalase (CAT) in HDL samples incubated with red blood cells (RBC) and lysed RBCs (RBClys). Quantification of densitometric analysis is shown (N = 3). **p < 0.01, ***p < 0.001.

iTRAQ HDL	Control (N = 7)	AAA (N = 14)	p-Value
Sex (male/female)	7/0	14/0	ns
Age (years ±SD)	64.9 ± 0.2	66.5 ± 1.2	0.024
Dyslipidemia (%)	64.9	64.9	ns
Current smoker (%)	71.4	50	ns
Diabetes (%)	14.3	14.3	ns
Hypertension (%)	42.9	64.3	ns
Heart Disease (%)	28.6	35.7	ns
Statins (%)	42.9	85.7	0.006
ELISA PRDX6	**Control (N = 27)**	**AAA (N = 47)**	**p-Value**
Sex (male/female)	27/0	47/0	ns
Age (years ±SD)	64.9 ± 0.2	66.1 ± 5.5	ns
Dyslipidaemia (%)	55.6	56.5	ns
Current smoker (%)	7.4	38.3	0.03
Diabetes (%)	7.4	19.1	ns
Hypertension (%)	25.9	63.8	0.01
Heart Disease (%)	18.5	17	ns
Statins (%)	11.1	38.3	ns

Table 1. Clinical characteristics of patients and control participants.

Discussion

HDLs are a complex and heterogeneous family of particles with different lipid and protein cargo and functionality. Different methodologies can be used to isolate HDLs, and not all HDL subpopulations are equivalent. In this study, we isolated HDLs by the well-established ultracentrifugation method to isolate most particle subclasses and obtain a general proteomics view of HDLs[21]. However, it is important to note that the small size of HDLs means that all the proteins detected and quantified by mass spectrometry are unlikely to reside in the same particle simultaneously. HDLs are currently regarded as a transport platform of constitutive and non-permanently associated lipids and proteins that may have specific functions in disease[17]. Our analysis identified and quantified 112 proteins that are consistently associated with HDLs in AAA, of which 34 proteins (8% of HDL by composition) have not been previously characterized as HDL components. Quantitative proteomics demonstrated that HDLs in AAA patients are particularly enriched in PRDX6, HLA-I, retinol-binding protein 4, and PON1 and are depleted in C4b-binding protein alpha chain and α-2 macroglobulin, among others. Systems biology analysis

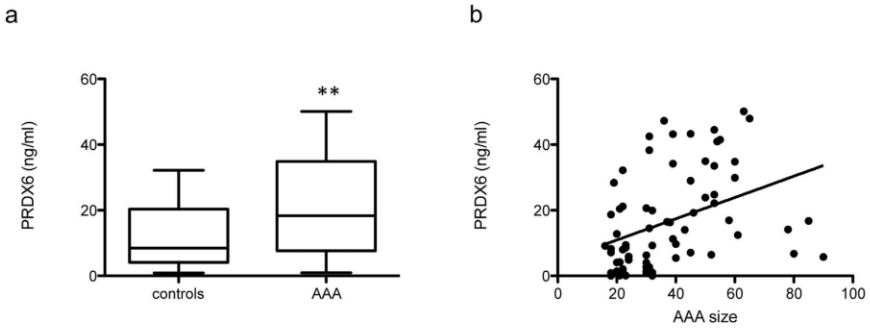


Figure 7. ELISA analysis of PRDX6 in AAA patient plasma. (**a**) Plasma concentration of PRDX6 in a cohort of control participants (N = 27) and AAA patients (N = 47)(**$p < 0.01$). (**b**) Correlation between PRDX6 and AAA size ($r = 0.4$, $p < 0.001$, age-adjusted).

demonstrated that these prominent changes reflect a general increase in proteins related to redox homeostasis, acute-phase response, and platelet activation and a decrease in proteins related to complement activation.

The levels and functionality of HDLs are altered in immune inflammatory diseases such as rheumatoid arthritis and systemic lupus erythematosus (SLE)[22]. Interestingly, serum amyloid A, one of the HDL components we found to be increased in AAA, has been suggested to underlie the impaired anti-inflammatory properties of HDL in SLE patients[23], and lack of endogenous acute-phase serum amyloid A protects against experimental AAA[24]. We also found that AAA patients have below-normal levels of C3, consistent with our previous finding that C3 declines in AAA patient plasma, contrasting its accumulation, consumption, and activation in the AAA thrombus[25]. In addition to the known role of HDLs in the humoral response through modulation of the complement system, HDL particles may also influence the innate and adaptive immune responses by modulating antigen presentation functions in macrophages, B cells, and T cells[26]. In line with this idea, we found that HDLs from AAA patients have increased levels of HLA-I, a central immune system molecule that presents endoplasmic-reticulum-derived peptides. The identification of HLA-I in HDL supports the role of HDL as a platform for the assembly of innate immune complexes[27]. Wang *et al.* demonstrated that dysfunctional HDLs promote lipid raft disruption, resulting in less control of antigen presentation and T cell activation[28]. Together, these results support an important role for HDLs in innate and adaptive immune responses, not only as a passive transport platform, but also through the active contribution of constituent proteins.

Proteins with roles in redox balance have been previously observed in HDLs[12]. Here, PON1 and PRDX6, two proteins implicated in oxidative stress, were increased in HDL from AAA patients. PON1 is known for its HDL-associated antioxidant capacity[29]. Vaisar *et al.*[12] reported elevated HDL PON1 content in patients with coronary artery disease, fitting well with our observation in AAA patients. However, we recently found that serum PON1 activity is decreased in AAA patients[30], suggesting that PON1 activity is impaired in HDLs of AAA patients. PRDX6 belongs to the peroxiredoxin family, a set of enzymes implicated in the protection against oxidation and the control of H_2O_2 signaling[31]. Other members of the PRDX family have been previously associated with AAA, such as PRDX-1 and PRDX-2[32,33]. However, this is the first time that PRDX6 has been associated with both HDL and AAA. PRDX6 is a bifunctional enzyme, with glutathione peroxidase and phospholipase A2 (PLA2) activities. Because of this dual function, the precise role of PRDX6 is not yet completely understood. Similar to other PRDXs, PRDX6 is able to reduce short-chain hydroperoxides through its peroxidase activity. However, the PLA2 activity specific to PRDX6 has been linked not only to antioxidant properties[34] but also to pro-oxidant properties[35]. Very recently, PRDX6 expression was shown to support higher Nox1-derived superoxide production, which was reduced by an inhibitor of PRDX6 phospholipase A2 activity[36]. The impaired antioxidant function reported in HDLs from AAA patients[19] will likely be the result of the imbalance between the levels and activities of the various pro- and antioxidant proteins, including PRDX6.

Previous studies of AAA tissue have shown increased levels of various proteins involved in redox balance. Elevated levels have been reported of the antioxidant proteins thioredoxin-1, PRDX1, and catalase in AAA thrombus, associated with both RBCs and neutrophils[37]. In the present study, high PRDX6 levels were also observed in AAA thrombus, mainly in areas of degenerated RBCs (in the process of erythrophagocytosis) and neutrophils. It is important to note that PRDX6 relocates to the cell membrane during neutrophil activation[38], which is required for optimal NADPH oxidase activity. The localization of PRDX6 in VSMCs of the AAA wall is probably a response to increased oxidative stress[39], since PRDX6 expression in VSMCs is known to increase in response to H_20_2[40]. Other oxidative stress markers identified in AAA tissue include the lipid peroxidation marker MDA and ceroids, which mark oxidation associated with lipids and RBCs[19]. We detected colocalization of PRDX6 with ceroids and MDA in the AAA thrombus; moreover, an intense PRDX6 signal was detected in cholesterol-rich and acellular atherosclerotic plaques in the AAA wall, colocalizing with MDA-positive areas. Cell-associated PRDX6 thus might participate in redox imbalance in AAA tissue through protective antioxidant functions or deleterious neutrophil-dependent NADPH activation; however, in a highly oxidative environment such as AAA, PRDX6 may lose its antioxidant activities through oxidative modification by lipid hydroperoxides[41]. Whether PRDX6 plays a protective or deleterious role in AAA tissue deserves further investigation.

Our finding that *ex-vivo* incubation of HDLs with lysed RBCs increases HDL levels of PRDX6 and catalase suggests that part of the PRDX6 observed in AAA thrombus may arise from RBC lysis. However, increased levels

of PRDX6, but not catalase, were also observed in HDLs incubated with intact RBCs, suggesting that HDLs may interact directly with membrane-bound PRDX6 in RBCs. In any case, PRDX6 translocation to the cell membrane is important for reactive oxygen species production through NADPH oxidase 2 complex activation[34,38], suggesting that the elevated PRDX6 levels in circulating HDL of AAA patients reflect an increased systemic oxidative stress. Our study not only identifies PRDX6 as a HDL constituent, but also shows that PRDX6 concentration doubles in the plasma of AAA patients, probably reflecting the systemic response to increased oxidative stress in AAA. We also found a positive correlation between PRDX6 and AAA size, a marker of AAA progression and the clinical parameter used in the management of AAA patients.

Our analysis suggests that the altered protein profile of HDLs in AAA reflects disease events, including an increase in antioxidant proteins probably associated with a systemic response to the redox imbalance in AAA. The increased HDL and plasma levels of PRDX6 in AAA patients support the potential of PRDX6 as a new biomarker of AAA.

Methods

The authors declare that all methods were performed in accordance with the relevant guidelines and regulations.

Patient selection. The studies were approved by the Research and Ethics Committee of the Fundación Jiménez Díaz University Hospital Health Research Institute (IIS-FJD; Madrid, Spain), and patients and control participants gave informed consent for their inclusion in the study. Patients with an asymptomatic infrarenal AAA (aortic size >3 cm confirmed by abdominal ultrasound) were recruited during clinical examination or before surgical repair at the Vascular Surgery Service at FJD University Hospital. Controls with non-dilated infrarenal aortas (aortic size <3 cm, confirmed by abdominal ultrasound) were recruited through a screening program. For proteomic analysis, EDTA plasma samples were obtained from 14 male AAA patients and 7 male control participants. For ELISA, a second set of plasma samples (27 controls and 47 AAA patients) were obtained from the IIS-FJD biobank. Clinical characteristics of all controls and AAA patients are summarized in Table 1. For immunohistochemistry, samples of AAA thrombus tissue (n = 10) and wall tissue (n = 10) were collected from male patients (70 ± 6 years old, 70% hypertensive, 30% current smokers, 50% dyslipidemic, 10% diabetic, 20% heart disease) undergoing open surgical repair due to aortic dilation >5 cm at the IIS-FJD Vascular Surgery Service.

HDL isolation. Lipoproteins were isolated from individual EDTA plasma samples by ultracentrifugation as described in Supplementary Information. Moreover, additional HDLs were isolated from healthy volunteers for incubation with red blood cells (RBC)(N = 3). Briefly, HDLs were incubated with RBCs (intact or lysed with H_2O/NaCl) for 4 hours at 37 °C, and HDLs were subsequently re-isolated by ultracentrifugation.

Proteomics. Proteomic analysis was performed on HDL particles isolated from 14 AAA patients and 7 controls (Table 1). HDL samples were in-gel digested overnight at 37 °C with sequencing-grade trypsin (Promega, Madison, WI, USA) at an 8:1 protein:trypsin (w/w) ratio in 50 mM ammonium bicarbonate, pH 8.8[42]. The resulting peptides were desalted on C18 Oasis cartridges (Waters Corporation, Milford, MA, USA) using 50% acetonitrile (ACN) (v/v) in 0.1% trifluoroacetic acid (v/v) as eluent, and vacuum dried. A total of 7 independent isobaric tags were performed for relative and absolute quantitation (iTRAQ) 4-plex experiments. iTRAQ labeling was performed essentially according to the manufacturer's instructions, as previously described in detail[42–44] (Supplementary Information). In each experiment, samples from 2 AAA patients and 1 control participant were compared with an internal control, prepared by pooling protein extracts from all subjects of the study. All the comparisons included independent biological preparations, making a total of 14 comparisons between AAA samples and the internal control sample and 7 comparisons between control samples and the internal control sample. The use of the internal control allowed comparison of data from different individuals in different experiments.

The tryptic peptide mixtures were subjected to nano-HPLC (Easy nLC 1000 liquid chromatograph, Thermo Scientific, San Jose, CA, USA) coupled to a Q Exactive mass spectrometer (Thermo Scientific). Peptides were suspended in 0.1% formic acid, loaded onto a C18 RP nano-precolumn (75 μm I.D. and 2 cm, Acclaim PepMap100, Thermo Scientific), and separated on an analytical C18 nano-column (75 μm I.D. and 50 cm, Acclaim PepMap100) in a continuous gradient increasing from 8% to 30% B over 120 min, followed by a rapid increase from 30% to 90% B over 2 min at a flow rate of 200 nL/min. The Q Exactive mass spectrometer was operated in data-dependent mode with a normal FT-resolution spectrum (70,000 resolution) in the mass range of m/z 390–1500, followed by acquisition of data-dependent MS/MS spectra from the 10 most intense parent ions identified in the chromatographic run.

Peptides were identified by searching against a Human Uniprot database supplemented with porcine trypsin (120501 entries; release October 2012). The search was conducted with the SEQUEST algorithm (Proteome Discoverer 1.4, Thermo Finnigan), allowing two missed cleavages and using 600 ppm precursor mass tolerance and 0.03 ppm fragment mass tolerance. Methionine oxidation and cysteine carbamidomethylation were allowed as variable modifications. For peptide iTRAQ labeling, lysine and N-terminal modifications of +144.1020 Da were selected as fixed modifications. The same MS/MS spectra collections were searched against inverted databases constructed from the same target databases. SEQUEST results were analyzed by the probability ratio method[45]; false discovery rates (FDR) for peptide identification were calculated using the refined method[46]. The statistical model used to analyze the quantitative data has been described before in detail[43] (Supplementary Information). The systems biology analysis was performed using the Systems Biology Triangle (SBT)[20].

The data set from the analysis of HDL proteome (raw and msf files, protein database fasta file, searching parameters xml file, and excel tables with identification and quantification data) is available in the PeptideAtlas repository (http://www.peptideatlas.org/PASS/PASS00861), which can be downloaded via ftp.peptideatlas.org.

Western blot. Equal amounts of HDL (20 μg) were loaded onto 10% polyacrylamide gels, electrophoresed and transferred to nitrocellulose membranes. Blots were blocked with 7% died skimmed milk in 0.05% Tris-buffered saline and Tween (TBS-T) for 1 hour and incubated overnight at 4 °C with the following antibodies: anti-PRDX6 (ab16947, abcam), anti-HLA-I (LS-B6775, LifeSpan Biosciences, Inc), anti-paraoxonase (PON1) (ab24261, abcam), anti-ApoA1 (home-made), or anti-catalase (ab52477, abcam). ApoA1 was detected as a loading control. Membranes were washed with TBS-T and incubated with the appropriate secondary antibody (1:2500) for 1 hour at room temperature. After 4 washes, the signal was detected with an ECL chemiluminiscence kit (GE Healthcare).

Histology and immunohistochemistry. Samples of arterial wall and intra-luminal thrombus obtained from AAA patients were embedded in paraffin, and 4 μm cross-sections were cut. Ceroids were detected by direct observation of tissue by fluorescence microscopy (ceroids autofluoresce at 550 nm, producing a red signal). Immunohistochemistry was performed with antibodies against the following proteins: PRDX6 (ab16947, abcam), the lipid peroxidation marker MDA (ab6463, abcam), the neutrophil marker CD15 (Dako), and alpha smooth muscle actin (Dako). Sections were then incubated with the appropriate biotinylated secondary antibody and ABComplex, followed by staining with 3,30-diaminobenzidine (DAB), hematoxylin counterstaining, and mounting in DPX medium.

ELISA. The plasma concentration of soluble PRDX6 in AAA and control samples was measured with a commercial ELISA kit (LF-EK0206, AbFrontier).

Statistical analysis. Data are expressed as mean ± SEM. Between-group comparisons were assessed for categorical variables with the X^2 test and for numerical variables by Mann-Whitney non-parametric test (ELISA and western blot of controls vs patients) or ANOVA followed by Bonferroni test (western blot of *in vitro* experiment). Multivariate logistic regression analysis included only variables that were statistically significant in the univariate analysis, and was performed to assess predictors of the presence of AAA. Univariate association of PRDX6 with AAA size was assessed by the Pearson correlation test and then adjusted for age. Ninety-five percent confidence intervals (CI) were calculated for each comparison. Differences were considered statistically significant at $p < 0.05$. Statistical analysis was performed with SPSS 15.0.

References

1. Jacomelli, J., Summers, L., Stevenson, A., Lees, T. & Earnshaw, J. J. Impact of the first 5 years of a national abdominal aortic aneurysm screening programme. *Br J Surg* **103,** 1125–1131, doi: 10.1002/bjs.10173 (2016).
2. Salvador-González, B. *et al.* Prevalence of Abdominal Aortic Aneurysm in Men Aged 65–74 Years in a Metropolitan Area in North-East Spain. *Eur J Vasc Endovasc Surg* **52,** 75–81, doi: 10.1016/j.ejvs.2016.04.005 (2016).
3. Saḳalihasan, N., Limet, R. & Defawe, O. D. Abdominal aortic aneurysm. *Lancet* **365,** 1577–1589, doi: 10.1016/S0140-6736(05)66459-8 (2005).
4. Limet, R., Saḳalihassan, N. & Albert, A. Determination of the expansion rate and incidence of rupture of abdominal aortic aneurysms. *J Vasc Surg* **14,** 540–548, doi: 10.1016/0741-5214(91)90249-T (1991).
5. Acosta-Martin, A. E. *et al.* Quantitative mass spectrometry analysis using PAcIFIC for the identification of plasma diagnostic biomarkers for abdominal aortic aneurysm. *PLoS One* **6,** e28698, doi: 10.1371/journal.pone.0028698 (2011).
6. Wallinder, J., Bergström, J. & Henriksson, A. E. Discovery of a novel circulating biomarker in patients with abdominal aortic aneurysm: a pilot study using a proteomic approach. *Clin Transl Sci* **5,** 56–59, doi: 10.1111/j.1752-8062.2011.00372.x (2012).
7. Spadaccio, C. *et al.* Serum proteomics in patients with diagnosis of abdominal aortic aneurysm. *Cardiovasc Pathol* **21,** 283–290, doi: 10.1016/j.carpath.2011.09.008 (2012).
8. Gamberi, T. *et al.* A proteomic approach to identify plasma proteins in patients with abdominal aortic aneurysm. *Mol Biosyst* **7,** 2855–2862, doi: 10.1039/c1mb05107e (2011).
9. Burillo, E. *et al.* ApoA-I/HDL-C levels are inversely associated with abdominal aortic aneurysm progression. *Thromb Haemost* **113,** 12, doi: 10.1160/TH14-10-0874 (2015).
10. Stather, P. W. *et al.* Meta-analysis and meta-regression analysis of biomarkers for abdominal aortic aneurysm. *Br J Surg* **101,** 1358–1372, doi: 10.1002/bjs.9593 (2014).
11. Jorge, I. *et al.* The human HDL proteome displays high inter-individual variability and is altered dynamically in response to angioplasty-induced atheroma plaque rupture. *J Proteomics* **106,** 61–73, doi: 10.1016/j.jprot.2014.04.010 (2014).
12. Vaisar, T. *et al.* Shotgun proteomics implicates protease inhibition and complement activation in the antiinflammatory properties of HDL. *J Clin Invest* **117,** 746–756, doi: 10.1172/JCI26206 (2007).
13. Alwaili, K. *et al. The* HDL proteome in acute coronary syndromes shifts to an inflammatory profile. *Biochim Biophys Acta* **1821,** 405–415, doi: 10.1016/j.bbalip.2011.07.013 (2012).
14. Tan, Y. *et al.* Acute coronary syndrome remodels the protein cargo and functions of high-density lipoprotein subfractions. *PLoS One* **9,** e94264, doi: 10.1371/journal.pone.0094264 (2014).
15. Yan, L. R. *et al.* A pro-atherogenic HDL profile in coronary heart disease patients: an iTRAQ labelling-based proteomic approach. *PLoS One* **9,** e98368, doi: 10.1371/journal.pone.0098368 (2014).
16. Annema, W. & von Eckardstein, A. High-density lipoproteins. Multifunctional but vulnerable protections from atherosclerosis. *Circ J* **77,** 2432–2448 doi: 10.1253/circj.CJ-13-1025 (2013).
17. Vickers, K. C. & Remaley, A. T. HDL and cholesterol: life after the divorce? *J Lipid Res* **55,** 4–12, doi: 10.1194/jlr.R035964 (2014).
18. Ortiz-Muñoz, G. *et al. HDL* antielastase activity prevents smooth muscle cell anoikis, a potential new antiatherogenic property. *FASEB J* **23,** 3129–3139, doi: 10.1096/fj.08-127928 (2009).
19. Delbosc, S. *et al.* Impaired high-density lipoprotein anti-oxidant capacity in human abdominal aortic aneurysm. *Cardiovasc Res* **100,** 307–315, doi: 10.1093/cvr/cvt194 (2013).
20. García-Marqués, F. *et al.* A novel systems-biology algorithm for the analysis of coordinated protein responses using quantitative proteomics. *Mol Cell Proteomics* **15,** 1740–1760, doi: 10.1074/mcp.M115.055905 (2016).
21. Rosenson, R. S. *et al.* HDL Measures, Particle Heterogeneity, Proposed Nomenclature, and Relation to Atherosclerotic Cardiovascular Events. *Clin Chem* **57,** 392–410, doi: 10.1373/clinchem.2010.155333 (2011).
22. McMahon, M. *et al.* Proinflammatory high-density lipoprotein as a biomarker for atherosclerosis in patients with systemic lupus erythematosus and rheumatoid arthritis. *Arthritis Rheum* **54,** 2541–2549, doi: 10.1002/art.21976 (2006).
23. Han, C. Y. *et al.* Serum amyloid A impairs the antiinflammatory properties of HDL. *J Clin Invest* **126,** 796, doi: 10.1172/JCI86401 (2016).

24. Webb, N. R. *et al.* Deficiency of Endogenous Acute-Phase Serum Amyloid A Protects apoE−/− Mice From Angiotensin II-Induced Abdominal Aortic Aneurysm Formation. *Arterioscler Thromb Vasc Biol* **35,** 1156–1165, doi: 10.1161/ATVBAHA.114.304776 (2015).
25. Martinez-Pinna, R. *et al.* Proteomic analysis of intraluminal thrombus highlights complement activation in human abdominal aortic aneurysms. *Arterioscler Thromb Vasc Biol* **33,** 2013–2020, doi: 10.1161/ATVBAHA.112.301191 (2013).
26. Norata, G. D., Pirillo, A., Ammirati, E. & Catapano, A. L. Emerging role of high density lipoproteins as a player in the immune system. *Atherosclerosis* **220,** 11–21, doi: 10.1016/j.atherosclerosis.2011.06.045 (2012).
27. Shiflett, A. M., Bishop, J. R., Pahwa, A. & Hajduk, S. L. Human high density lipoproteins are platforms for the assembly of multi-component innate immune complexes. *J Biol Chem* **280,** 32578–32585, doi: 10.1074/jbc.M503510200 (2005).
28. Wang, S. H., Yuan, S. G., Peng, D. Q. & Zhao, S. P. HDL and ApoA-I inhibit antigen presentation-mediated T cell activation by disrupting lipid rafts in antigen presenting cells. *Atherosclerosis* **225,** 105–114, doi: 10.1016/j.atherosclerosis.2012.07.029 (2012).
29. Soran, H., Schofield, J. D., Liu, Y. & Durrington, P. N. How HDL protects LDL against atherogenic modification: paraoxonase 1 and other dramatis personae. *Curr Opin Lipidol* **26,** 247–256, doi: 10.1097/MOL.0000000000000194 (2015).
30. Burillo, E. *et al.* Paraoxonase-1 overexpression prevents experimental abdominal aortic aneurysm progression. *Clin Sci* (*Lond*) **130,** 10, doi: 10.1042/CS20160185 (2016).
31. Ambruso, D. R. Peroxiredoxin-6 and NADPH oxidase activity. *Methods Enzymol* **527,** 145–167, doi: 10.1016/B978-0-12-405882-8.00008-8 (2013).
32. Martinez-Pinna, R. *et al.* Identification of peroxiredoxin-1 as a novel biomarker of abdominal aortic aneurysm. *Arterioscler Thromb Vasc Biol* **31,** 935–943, doi: 10.1161/ATVBAHA.110.214429 (2011).
33. Martinez-Pinna, R., Gonzalez de Peredo, A., Monsarrat, B., Burlet-Schiltz, O. & Martin-Ventura, J. L. Label-free quantitative proteomic analysis of human plasma-derived microvesicles to find protein signatures of abdominal aortic aneurysms. *Proteomics Clin Appl* **8,** 620–625, doi: 10.1002/prca.201400010 (2014).
34. Lien, Y. C., Feinstein, S. I., Dodia, C. & Fisher, A. B. The roles of peroxidase and phospholipase A2 activities of peroxiredoxin 6 in protecting pulmonary microvascular endothelial cells against peroxidative stress. *Antioxid Redox Signal* **16,** 440–451, doi: 10.1089/ars.2011.3950 (2012).
35. Chatterjee, S. *et al.* Peroxiredoxin 6 phosphorylation and subsequent phospholipase A2 activity are required for agonist-mediated activation of NADPH oxidase in mouse pulmonary microvascular endothelium and alveolar macrophages. *J Biol Chem* **286,** 11696–11706, doi: 10.1074/jbc.M110.206623 (2011).
36. Kwon, J. *et al.* Peroxiredoxin 6 (Prdx6) supports NADPH oxidase1 (Nox1)-based superoxide generation and cell migration. *Free Radic Biol Med* **96,** 99–115, doi: 10.1016/j.freeradbiomed.2016.04.009 (2016).
37. Martin-Ventura, J. L. *et al.* Erythrocytes, leukocytes and platelets as a source of oxidative stress in chronic vascular diseases: detoxifying mechanisms and potential therapeutic options. *Thromb Haemost* **108,** 435–442, doi: 10.1160/TH12-04-0248 (2012).
38. Ambruso, D. R., Ellison, M. A., Thurman, G. W. & Leto, T. L. Peroxiredoxin 6 translocates to the plasma membrane during neutrophil activation and is required for optimal NADPH oxidase activity. *Biochim Biophys Acta* **1823,** 306–315, doi: 10.1016/j.bbamcr.2011.11.014 (2012).
39. Chowdhury, I. *et al.* Oxidant stress stimulates expression of the human peroxiredoxin 6 gene by a transcriptional mechanism involving an antioxidant response element. *Free Radic Biol Med* **46,** 146–153, doi: 10.1016/j.freeradbiomed.2008.09.027 (2009).
40. Lee, C. K. *et al.* Analysis of peroxiredoxin decreasing oxidative stress in hypertensive aortic smooth muscle. *Biochim Biophys Acta* **1774,** 848–855, doi: 10.1016/j.bbapap.2007.04.018 (2007).
41. Zarkovic, N., Cipak, A., Jaganjac, M., Borovic, S. & Zarkovic, K. Pathophysiological relevance of aldehydic protein modifications. *J Proteomics* **92,** 239–247, doi: 10.1016/j.jprot.2013.02.004 (2013).
42. Bonzon-Kulichenko, E. *et al.* A robust method for quantitative high-throughput analysis of proteomes by 18O labeling. *Mol Cell Proteomics* **10,** M110 003335, doi: 10.1074/mcp.M110.003335 (2011).
43. Navarro, P. *et al.* General statistical framework for quantitative proteomics by stable isotope labeling. *J Proteome Res* **13,** 1234–1247, doi: 10.1021/pr4006958 (2014).
44. Martinez-Acedo, P. *et al.* A novel strategy for global analysis of the dynamic thiol redox proteome. *Molecular & cellular proteomics: MCP* **11,** 800–813, doi: 10.1074/mcp.M111.016469 (2012).
45. Martínez-Bartolomé, S. *et al.* Properties of average score distributions of SEQUEST: the probability ratio method. *Mol Cell Proteomics* **7,** 1135–1145, doi: 10.1074/mcp.M700239-MCP200 (2008).
46. Navarro, P. & Vázquez, J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *J Proteome Res* **8,** 1792–1796, doi: 10.1021/pr800362h (2009).

Acknowledgements

We thank Simon Bartlett for language and scientific editing. This study was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) (SAF2016-80843-R, BIO2012–37926 and BIO2015-67580-P), Fondo de Investigaciones Sanitarias ISCiii-FEDER (PRB2) (IPT13/0001, ProteoRed, Redes RIC RD12/0042/00038 and RD12/0042/0056, Biobancos RD09/0076/00101 and CA12/00371), Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), and FRIAT. The CNIC is supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the Pro-CNIC Foundation, and is a Severo Ochoa Center of Excellence (MINECO award SEV-2015-0505).

Author Contributions

E.B. and I.J. performed all the experiments, prepared the figures and wrote the manuscript. D.M. and E.C. helped in the proteomic analysis and in the *in vitro* experiments. M.T.H. and I.E. designed and performed the quantification and statistical proteomic method. L.B.C., J.E., J.B.M. and O.M. contributed with scientific support. J.V. and J.L.M.V. designed and supervised the experiments, correct the manuscript and coordinated the project. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Burillo, E. *et al.* Quantitative HDL Proteomics Identifies Peroxiredoxin-6 as a Biomarker of Human Abdominal Aortic Aneurysm. *Sci. Rep.* **6**, 38477; doi: 10.1038/srep38477 (2016).

Appendix 2: Help of the programs in the SanXoT software package

2.1 Aljamia

```
Aljamia v1.09 is a program made in the Jesus Vazquez Cardiovascular
Proteomics Lab at Centro Nacional de Investigaciones Cardiovasculares, used to
convert data in xml tables into a tab-separated values archive.

Aljamia needs an XML input file, and:

     * up to four strings to combine information from the xml field.
       Commands: -i, -j, -k and -l. Usage: -i[FirstScan] -j[Sequence]
       It is possible to combine fields: -i[RAWFileName]-[FirstScan]_[Charge]
       (which would deliver something like "sampleA.raw-1029-3").
       Everything outside brackets will be copied unchanged. Note that the
       fields are case sensitive.
     * the name of the table where these fields are (command -t). Default is
       "peptide_match".

And delivers:

     * an output data file with three columns (id, x, v) suitable to work as
       input for SanXoT.

Usage: aljamia.py -x[xml file] -i[fold field] [-j[weight field] -k[id string], ...]
[OPTIONS]

   -h, --help          Display this help and exit.
   -a, --analysis=string
                       Use a prefix for the output files. If this is not
                       provided, then the prefix will be garnered from the data
                       file.
   -A, --allow-operations
                       Allow python-style operations for the indicated fields.
                       Example: having Scan = 900, Charge = 3, and using
                           -i"[Scan]-[Charge]" -j"[Scan]-[Charge]" -A"i"
                       Will return "887" and "900-3" i- and j-fields,
                       respectively. By default, no operations are allowed.
   -d, --allow-duplicates
                       To avoid removal of duplicated relations.
   -f, --filter=string To filter data to import. Use as in these examples:

                           -f"[Charge]==2"
                           -f"[st_excluded]!=excluded", which means
                               st_excluded must NOT be equal to "excluded"
                           -f"[Charge]=2&&[st_excluded]!=excluded", which
                               means charge must be 2, and st_excluded must
                               not be equal to "excluded"
                           -f"[FirstScan]>=1000"
                           -f"[FASTAProteinDescription]~~clathrin", which means
                                FASTAProteinDescription must include "clathrin"
                           -f"[Sequence]!~C", which means
                                Sequence must NOT include "C"
                           -f"[Sequence]!=ABABABABK", which means
                                Sequence must be different than "ABABABABK"
                           -f"!([Sequence]~~C || [Sequence]~~M)", which means
                                Sequence must not (via "!") contain "C" or
                                (via "||") "M". Note you can use parentheses
                           -f"[Sequence]~~C && [Sequence]~~M", wchich means
                                Sequence must contain "C" and (via "&&") "M"

                       Note that the filter is case sensitive.
                       In forthcoming versions filters will be available for
                       numerical operations, but currently the filter doesn't
                       work with conditionals such as [Mass] > 565.2, only for
                       (in)equalities such as [Mass] == 565.2

   -i, --id1=string    Identifier for the first column. XML tags must be in
                       square brackets, while the rest of the text will be kept
                       unaltered. Here are some examples using tags such as
                       "FirstScan", "Charge", "Mass" or "Sequence" or "PTM":
```

```
                        "ABCD" -> "ABCD" (no tags -> unchanged, to all rows)
                        "FS[FirstScan]_q=[Charge]" -> "FS2991_q=2"
                        "ABCD-[Charge]" -> "ABCD-3"
                        "ABCD_[Charge]_[Mass]" -> "ABCD_3_578.1684"
                        "[Sequence]_[PTM]" -> "SAPEREAVDEK_15.994915"

                    Note that tags are case-sensitive.

  -j, --id2=string    Identifier for the second column (see -i).
  -k, --id3=string    Identifier for the third column (see -i).
  -l, --id4=string    Identifier for the fourth column (see -i).
  -L, --logfile=filename
                    To use a non-default name for the log file.
  -o, --output=filename
                    To use a non-default name for the output file.
  -p, --place, --folder=foldername
                    To use a different common folder for the output files.
                    If this is not provided, the the folder used will be the
                    same as the input folder.
  -R, --initialrow=integer
                    To set the position of row with headers (default is 1).
  -t, --table=number  To select fields from a table different than QuiXML's
                    peptide_match (which corresponds to the default, 3).
  -x, --input=filename, --filename=filename
                    Input xml or txt (tsv) file.
```

2.2 Anselmo

```
Anselmo v0.04 is a program made in the Jesus Vazquez Cardiovascular Proteomics
Lab at Centro Nacional de Investigaciones Cardiovasculares, used to identify
the integration which holds the median variance from a set of randomised
SanXoT integrations (i.e., integrations performed using SanXoT's parametre -R).

Anselmo needs at least:

1) the prefix (by using the -f argument) of the set of experiments, i.e. when
the info files of the randomisations are such as:

    heartExperiment1_infoFile.txt
    heartExperiment2_infoFile.txt
    ...
    heartExperimentZ_infoFile.txt

the prefix is considered to be "heartExperiment".

2) the folder where all those info files are (they must be in the same folder),
using the -p argument.

After reading the variances in the set of info files, Anselmo identifies the
info file containing the median of the set of variances, and then it renames
the files using its prefix, i.e. using the previous example, and assuming the
median of the variance is in experiment labelled "9", then it copies the files

    heartExperiment9_infoFile.txt
    heartExperiment9_higherLevel.txt
    heartExperiment9_outStats.txt
    ...

into

    med_heartExperiment_infoFile.txt
    med_heartExperiment_higherLevel.txt
    med_heartExperiment_outStats.txt
    ...

Usage: anselmo.py -p[folder] -fyeast_nullHypothesis [OPTIONS]

Use -H or --advanced-help for more details.
```

2.3 Arbor

```
Arbor v1.04 is a program made in the Jesus Vazquez Cardiovascular
Proteomics Lab at Centro Nacional de Investigaciones Cardiovasculares, used to
generate the tree graph of a set of categories, showing the position of a given
list of categories in the tree, along with category information.

Arbor needs four input files:

     * a stats file, the outStats file from SanXoT (using the -z command); if
          this is omitted, then the tree will only distinguish the categories
         in the list from the other categories above them (not showing the
         values of the protein within the category).
     * a higher level list to graph (using the -c command)
     * a relations file (using -r command)
     * a list of links between higher level elements, such as the
         table_allPaths.xls from GOconnect (using the -b command)

And delivers three output files:

     * the graph in PNG format (default suffix: "_outTree.png")
     * the DOT language text file used to generate the graph (default suffix:
          "_outTree.gv")
     * a log file (default suffix: "_logFile")

Usage: arbor.py -z[stats file] -r[relations file] -c[higher level list file] -b[links
file] [OPTIONS]

   -h, --help          Display this help and exit.
   -a, --analysis=string
                       Use a prefix for the output files. If this is not
                       provided, then the prefix will be garnered from the
                       stats file.
   -b, --biglist       A list of links between higher level elements, such as
                       the table_allPaths.xls from GOconnect. It must be a tab
                       separated values text file, containing any identifier
                       in the first column (this column will not be imported,
                       but originally it was intended to contain protein
                       identifiers for each path), containing in each row (from
                       the second column on) a possible path from top to the
                       most specific element.
   -c, --list=filename The text file containing the higher level elements whose
                       categories we want to relate. If the first element is
                       not taken, it might help saving the file with ANSI
                       format. If a header is used, then it must be in the form
                       "id>n>Z>FDR" or "id>Z>n" (where ">" means "tab").
   -d, --dotfile=filename
                       To use a non-default name for the text file in DOT
                       language, which is used to generate the graph.
   -g, --graphformat=string
                       File format for the similarity graph (default is "png").
   -G, --outgraph=filename
                       To use a non-default name for the graph file.
   -l, --graphlimits=integer
                       To set the +- limits of the most intense red/green
                       colours in the graph (default is 6).
   -L, --logfile=filename
                       To use a non-default name for the log file.
   -N, --altmax=integer
                       Maximum number of lower level elements that the alt text
                       of the higher level node will show per side. For
                       instance, for N = 3, alt text will show all the elements
                       up to six; beyond this, only the first and last three
                       will be shown. (Default is N = 5.) (Note that this will
                       have effect if the SVG format is used.)
   -p, --place, --folder=foldername
                       To use a different common folder for the output files.
                       If this is not provided, the the folder used will be the
                       same as the stats file folder.
   -r, --relfile, --relationsfile=filename
                       Relations file, with identificators of the higher level
                       in the first column, and identificators of the lower
                       level in the second column.
   -z, --outstats=filename
                       The outStats file from a SanXoT integration (optional,
                       see above).
```

```
   --selectednodecolor=#rrggbb, --selectednodecolour=#rrggbb
   --defaultnodecolor=#rrggbb, --defaultnodecolour=#rrggbb
   --errornodecolor=#rrggbb, --errornodecolour=#rrggbb
   --mincolor=#rrggbb, --mincolour=#rrggbb
   --middlecolor=#rrggbb, --middlecolour=#rrggbb
   --maxcolor=#rrggbb, --maxcolour=#rrggbb
```

2.4 Cardenio

```
usage: cardenio.exe [-h] -a ANALYSIS -p PLACE [-L LOGFILE] -t TAGFILE
                    [-d DATAFILE] [-r RELFILE] [-s SEPARATOR] [-v]

Cardenio v0.03 is a program made in the Jesus Vazquez Cardiovascular
Proteomics Lab at Centro Nacional de Investigaciones Cardiovasculares, used to
generate joined integration data and relations files from different
experiments..

optional arguments:
  -h, --help            show this help message and exit
  -a ANALYSIS, --analysis ANALYSIS
                        Use a prefix for the output files.
  -p PLACE, --place PLACE
                        To use a different common folder for the output files.
                        If this is not provided, the the folder used will be
                        the same as the FASTA file folder.
  -L LOGFILE, --logfile LOGFILE
                        To use a non-default name for the log file.
  -t TAGFILE, --tagfile TAGFILE
                        The file containing the tags used for the different
                        experiments to be joined.
  -d DATAFILE, --datafile DATAFILE
                        To use a non-default merged data file name.
  -r RELFILE, --relfile RELFILE
                        To use a non-default merged relations file name.
  -s SEPARATOR, --separator SEPARATOR
                        To use a non-default suffix separator (default is
                        "_").
  -v, --verbose         To write down extra information about operations
                        performed.
```

2.5 Catapep

```
usage: catapep.exe [-h] -a ANALYSIS -p PLACE [-L LOGFILE] -d INPUTFILE -M
                   MSFFILE [-r RAWFILECOL] [-s SCANNUMBERCOL] [-q CHARGECOL]
                   [-S PEPSEQUENCECOL] [-x XCORRCOL] [-R INITIALROW] [-v] [-Q]
                   [-O]

cataPep v1.03 is a program made in the Jesus Vazquez Cardiovascular Proteomics
Lab at Centro Nacional de Investigaciones Cardiovasculares, used to make zero
the XCorrs of all the PSMs in an MSF file, excluding those provided in a
list..

optional arguments:
  -h, --help            show this help message and exit
  -a ANALYSIS, --analysis ANALYSIS
                        Use a prefix for the output files.
  -p PLACE, --place PLACE
                        To use a different common folder for the output files.
                        If this is not provided, the the folder used will be
                        the same as the FASTA file folder.
  -L LOGFILE, --logfile LOGFILE
                        To use a non-default name for the log file.
  -d INPUTFILE, --inputfile INPUTFILE
                        Name of the text file containing the list of PSMs to
                        keep in the MSF.
  -M MSFFILE, --msffile MSFFILE
                        Name of the MSF file having the PSMs to modify.
  -r RAWFILECOL, --rawfilecol RAWFILECOL
                        Header of the column contaning the name of the RAW
                        files. Default is "RAWFileName".
  -s SCANNUMBERCOL, --scannumbercol SCANNUMBERCOL
                        Header of the column containing the scan numbers.
```

```
                        Default is "FirstScan".
  -q CHARGECOL, --chargecol CHARGECOL
                        Header of the column containing the charge. Default is
                        "Charge".
  -S PEPSEQUENCECOL, --pepsequencecol PEPSEQUENCECOL
                        Header of the column containing the identified peptide
                        sequences. Default is "Sequence"
  -x XCORRCOL, --xcorrcol XCORRCOL
                        Header of the column containing the XCorr. Default is
                        "XC1D".
  -R INITIALROW, --initialrow INITIALROW
                        The position of the row containing the headers.
                        Default is 1.
  -v, --verbose         Show extra info while operating.
  -Q, --quixml          Use column headers for QuiXML results tab separated
                        table file (otherwise, pRatio results file headers
                        will be used by default).
  -O, --changeoriginalmsf
                        Do not copy the MSF file to be modified, just remove
                        bad PSMs in the original file.
```

2.6 Klibrate

```
Klibrate v1.14 is a program made in the Jesus Vazquez Cardiovascular Proteomics
Lab at Centro Nacional de Investigaciones Cardiovasculares, used to perform the
calibration of experimental data, as a first step to integrate these data into
higher levels along with the SanXoT program.

To perform the calibration two parameters have to be calculated: the k (weight
constant), and the variance. They are calculated iteratively using the
Levenberg-Marquardt algorithm, starting from the seeds the user introduces
(it is possible to perform the calculation without the iterative calculation by
forcing both parameters with the -f option). In the integration that follows
the variance can be recalculated.

Klibrate needs two input files:

     * the original data file, containing unique identifiers of each scan, such
     as "RawFile05.raw-scan19289-charge2" or "File05B_scannumber12877_z3", the
     Xi which corresponds to the log2(A/B), and the Vi which corresponds to the
     weight of the measure).

     * the relations file, containing a first column with the higher level
     identifiers (such as the peptide sequence, for example "CGLAGCGLLK", or
     the protein, if you wish to directly integrate scans into proteins, such
     as the Uniprot Accession Numbers "P01308" or KEGG Gene ID "hsa:3630"), and
     the lower level identifiers within the abovementioned original data file
     (such as "RawFile05.raw-scan19289-charge2").

And delivers the output calibrated file:

     * the calibrated data file, containing the same information as the
     original data file, but changing the values of the third column
     (containing the weights) to adapt the information to the calibrated
     weights that can be used as input in the SanXoT program.



Usage: klibrate.py [OPTIONS] -r[relations file] -d[original data file] -o[calibrated
output file]

   -h, --help          Display this help and exit.
   -a, --analysis=string
                       Use a prefix for the output files. If this is not
                       provided, then the prefix will be garnered from the data
                       file.
   -b, --no-verbose       Do not print result summary after executing.
   -d, --datafile      Input data file with text identificators in the first
                       column, measured values (x) in the second column, and
                       uncalibrated weights (v) in the third column.
   -D, --outgraphdata=filename
                       To use a non-default name for the data used to create
                       calibration graph files.
   -f, --forceparameters
                       Use the parameters (k and variance) as provided, without
                       using the Levenberg-Marquardt algorithm.
```

```
   -g, --no-showgraph  Do not show the rank(V) vs 1 / MSD graph after
                       the calculation.
   -G, --outgraphvvalue=filename
                       To use a non-default name for the graph file which shows
                       the value of V (the weight) versus 1 / MSD.
   -k, --kseed         Seed for the weight constant. Default is k = 1.
   -K, --kfile=filename
                       Get the K value from a text file. It must contain a line
                       (not more than once) with the text "K = [float]". This
                       suits the info file from another integration (see -L).
   -L, --infofile=filename
                       To use a non-default name for the info file.
   -m, --maxiterations Maximum number of iterations performed by the Levenberg-
                       Marquardt algorithm to calculate the variance and the k
                       constant. If unused, the default value of the algorithm
                       is taken.
   -o, --outputfile    To use a non-default output calibrated file name (see
                       above for more information on this file).
   -p, --place, --folder=foldername
                       To use a different common folder for the output files.
                       If this is not provided, the the folder used will be the
                       same as the input folder.
   -r, --relfile, --relationsfile
                       Relations file, with identificators of the higher level
                       in the first column, and identificators of the lower
                       level in the second column.
   -R, --outgraphvrank=filename
                       To use a non-default name for the graph file which shows
                       the rank of V (the weight) versus 1 / MSD.
   -s, --no-showsteps  Do not print result summary and steps of each Levenberg-
                       Marquardt iteration.
   -v, --var, --varianceseed
                       Seed used to start calculating the variance.
                       Default is 0.001.
   -V, --varfile=filename
                       Get the variance value from a text file. It must contain
                       a line (not more than once) with the text
                       "Variance = [double]". This suits the info file from
                       another integration (see -L).
   -w, --window        The amount of weight-ordered lower level elements
                       (scans, usually) that are taken at a time to calculate
                       the median of the weight, which is compared to the fit;
                       default is 200.


examples:

* To calculate the variance and k starting with a seed v = 0.03 and k = 40, printing the
steps of the Levenberg-Marquardt algorithm and results, showing the rank(Vs) vs 1 / MSD
graph afterwards:

klibrate.py -gbs -v0.03 -k40 -dC:\temp\originalDataFile.txt -rC:\temp\relationsFile.txt
-oC:\temp\calibratedWeights.xls

* To get fast results of an integration forcing a variance = 0.02922 and a k = 35.28:

klibrate.py      -f      -v0.02922      -k35.28      -dC:\temp\originalDataFile.txt      -
rC:\temp\relationsFile.txt -oC:\temp\calibratedWeights.xls

* To see the graph resulting from a calculation with variance = 0.02922 and a k = 35.28:

klibrate.py      -gf      -v0.02922      -k35.28      -dC:\temp\originalDataFile.txt      -
rC:\temp\relationsFile.txt -oC:\temp\calibratedWeights.xls
```

2.7 MaesePedro

```
MaesePedro v0.03 is a program made in the Jesus Vazquez Cardiovascular
Proteomics Lab at Centro Nacional de Investigaciones Cardiovasculares, used to
generate pseudoinverted FASTA files. In collaboration with Luis Ferrández.

Sanson needs one input files:

     * a FASTA file (using the -f command)

And delivers two output files:
```

```
     * the inverted FASTA file
     * a log file (default suffix: "_logFile")

Usage: maesepedro.py -f[FASTAFile]

   -h, --help          Display this help and exit.
   -a, --analysis=string
                       Use a prefix for the output files.
   -p, --place, --folder=foldername
                       To use a different common folder for the output files.
                       If this is not provided, the the folder used will be the
                       same as the FASTA file folder.
   -c, --cleavesites=string
                       The residues after which the protease is cleaving.
                       Default is trypsin (KR). Note that only C-terminal
                       cleaving proteases are being considered in this version.
   -f, --fastafile=string
                       The input FASTA file to invert.
   -r, --removepalindromes
                       Remove peptides unchanged upon pseudoinversion, i.e.,
                       peptides such as ASSAK, EGTGER (when using trypsin).
                       Palindromic peptides are not removed by default.
```

2.8 Sanson

```
Sanson v1.06 is a program made in the Jesus Vazquez Cardiovascular
Proteomics Lab at Centro Nacional de Investigaciones Cardiovasculares, used to
generate the similarity graph of a set of categories.

A similarity graph is a graph that shows the relationship between a set of
categories by taking into account how many proteins they share. This is
measured with a variable f such that for categories c1 and c2, we get:

   f(c1, c2) = (#proteins shared by c1 and c2) / (#proteins of c1)

for instance:
  * if c1 == c2, we get f(c1, c2) = f(c2, c1) = 1;
  * if c1 and c2 do not share any proteins, we get f(c1, c2) = f(c2, c1) = 0;
  * if c2 is contained in c1, we get f(c1, c2) <= 1, f(c2, c1) = 1, etc

If no f number is given with the parametres (-e), then the program
automatically calculates the best f number, by maximising both the number of
category clusters and the number categories within each cluster.

Sanson needs three input files:

     * a stats file, the outStats file from SanXoT (using the -z command)
     * a higher level list to graph (using the -c command)
     * a relations file (using -r command)

And delivers five output files:

     * the graph in PNG format (default suffix: "_simGraph.png")
     * the DOT language text file used to generate the graph (default suffix:
          "_simGraph.gv")
     * a table showing the clusters generated (default suffix: "_outClusters")
     * the similarity matrix used to generate the graph (default suffix:
          "_outSimilarities")
     * a log file (default suffix: "_logFile")

Usage: sanson.py -z[stats file] -r[relations file] -c[higher level list file] [OPTIONS]

   -h, --help          Display this help and exit.
   -a, --analysis=string
                       Use a prefix for the output files. If this is not
                       provided, then the prefix will be garnered from the
                       stats file.
   -b, --nosubstats    To avoid colouring the boxes according to the proteins
                       that are in the concerning category (in this case, the
                       box is coloured using the Zij of the category, when this
                       information is available in the higher level list to
                       graph, see -c command).
   -c, --list=filename The text file containing the higher level elements whose
                       categories we want to relate. If the first element is
                       not taken, it might help saving the file with ANSI
```

```
                    format. If a header is used, then it must be in the form
                    "id>n>Z>FDR" or "id>Z>n" (where ">" means "tab").
   -d, --dotfile=filename
                    To use a non-default name for the text file in DOT
                    language, which is used to generate the graph.
   -e, --similarity=float
                    To override the calculation of the optimal f number (see
                    above for more details).
   -g, --graphformat=string
                    File format for the similarity graph (default is "png").
   -G, --outgraph=filename
                    To use a non-default name for the graph file.
   -l, --graphlimits=integer
                    To set the +- limits of the most intense red/green
                    colours in the graph (default is 6).
   -L, --logfile=filename
                    To use a non-default name for the log file.
   -m, --simfile=string
                    To use a non-default name for the similarity matrix
                    file.
   -N, --altmax=integer
                    Maximum number of lower level elements that the alt text
                    of the higher level node will show per side. For
                    instance, for N = 3, alt text will show all the elements
                    up to six; beyond this, only the first and last three
                    will be shown. (Default is N = 5.) (Note that this will
                    have effect if the SVG format is used.)
   -p, --place, --folder=foldername
                    To use a different common folder for the output files.
                    If this is not provided, the the folder used will be the
                    same as the stats file folder.
   -r, --relfile, --relationsfile=filename
                    Relations file, with identificators of the higher level
                    in the first column, and identificators of the lower
                    level in the second column.
   -s, --outcluster=filename
                    To use a non-default name for the file containg the
                    list of clusters.
   -z, --outstats=filename
                    The outStats file from a SanXoT integration.

   --selectednodecolor=#rrggbb, --selectednodecolour=#rrggbb
   --defaultnodecolor=#rrggbb, --defaultnodecolour=#rrggbb
   --errornodecolor=#rrggbb, --errornodecolour=#rrggbb
   --mincolor=#rrggbb, --mincolour=#rrggbb
   --middlecolor=#rrggbb, --middlecolour=#rrggbb
   --maxcolor=#rrggbb, --maxcolour=#rrggbb
```

2.9 SanXoT

```
SanXoT v2.07 is a program made in the Jesus Vazquez Cardiovascular Proteomics
Lab at Centro Nacional de Investigaciones Cardiovasculares, used to perform
integration of experimental data to a higher level (such as integration from
peptide data to protein data), while determining the variance between them.

SanXoT needs two input files:

     * the lower level input data file, a tab separated text file containing
     three columns: the first one with the unique identifiers of each lower
     level element (such as "RawFile05.raw-scan19289-charge2" for a scan, or
     "CGLAGCGLLK" for a peptide sequence, or "P01308" for the Uniprot accession
     number of a protein), the Xi which corresponds to the log2(A/B), and the
     Vi which corresponds to the weight of the measure). This data have to be
     pre-calibrated with a certain weight (see help of the Klibrate program).

     * the relations file, a tab separated text file containing a first column
     with the higher level identifiers (such as the peptide sequence, a Uniprot
     accession number, or a Gene Ontology category) and the lower level
     identifiers within the abovementioned input data file.

     NOTE: you must include a first line header in all your files.

And delivers six output file:

     * the output data file for the higher level, which has the same format as
```

```
    the lower level data file, but containing the ids of the higher level in
    the first column, the ratio Xj in the second column, and the weight Vj in
    the third column. By default, this file is suffixed as "_higherLevel".

    * two lower level output files, containing three columns each: in both,
    the first column contains with the identifiers of the lower level, the
    second column contains the Xinf - Xsup (i.e. the ratios of the lower
    level, but centered for each element they belong to), and the third column
    is either the new weight Winf (contanining the variance of the
    integration) or the former untouched Vinf weight. For example, integrating
    from scan to peptide, these files would contain firstly the scan
    identifiers, secondly the Xscan - Xpep (the ratios of each scan compared
    to the peptide they are identifying) and either Wscan (the weight of the
    scan, taking into account the variance of the scan distribution) or Vscan.
    By default, these files are suffixed "_lowerNormW" and "_lowerNormV".

    * a file useful for statistics, containing all the relations of the higher
    and lower level element present in the data file, with a copy of their
    ratios X and weights V, followed by the number of lower elements contained
    in the upper element (for example, the number of scans that identify the
    same peptide), the Z (which is the distance in sigmas of the lower level
    ratio X to the higher level weighted average), and the FDR (the false
    discovery rate, important to keep track of changes or outliers). By
    default, this file is suffixed "_outStats".

    * an info file, containing a log of the performed integrations. Its last
    line is always in the form of "Variance = [double]". This file can be used
    as input in place of the variance (see -v and -V arguments). By default,
    this file is suffixed "_infoFile".

    * a graph file, depicting the sigmoid of the Z column which appears in the
    stats file, compared to the theoretical normal distribution. By default,
    this file is suffixed "_outGraph".

Usage: sanxot.py -d[data file] -r[relations file] [OPTIONS]

   -h, --help          Display basic help and exit.
   -H, --advanced-help Display this help and exit.
   -A, --infofile=filename
                       To use a non-default name for the randomised relations
                       file (only applicable when -R is in use).
   -a, --analysis=string
                       Use a prefix for the output files. If this is not
                       provided, then the prefix will be garnered from the data
                       file.
   -b, --no-verbose    Do not print result summary after executing.
   -C, --confluence    A modified version of the relations file is used, where
                       all the destination higher level elements are "1". If no
                       relations file is provided, the program gets the lower
                       level elements from the first column of the data file.
   -d, --datafile=filename
                       Data file with identificators of the lowel level in the
                       first column, measured values (x) in the second column,
                       and weights (v) in the third column.
   -D, --removeduplicateupper
                       When merging data with relations table, remove duplicate
                       higher level elements (not removed by default).
   -f, --forceparameters
                       Use the parameters as provided, without using the
                       Levenberg-Marquardt algorithm. Negative variances will
                       be reset to zero (see -F if you do not wish this).
   -F, --forcenegativevariance
                       Same effect as -f, but not correcting negative variance
                       (as using -f, the program will automatically assign
                       variance = 0 if forced variance is negative).
   -g, --no-graph      Do not show the Zij vs rank / N graph.
   -G, --outgraph=filename
                       To use a non-default name for the graph file.
   -l, --graphlimits=integer
                       To set the +- limits of the Zij graph (default is 6). If
                       you want the limits to be between the minimum and
                       maximum values, you can use -l.
   -L, --infofile=filename
                       To use a non-default name for the info file.
   -m, --maxiterations=integer
                       Maximum number of iterations performed by the Levenberg-
                       Marquardt algorithm to calculate the variance. If
                       unused, then the default value of the algorithm is
                       taken.
```

```
   -M, --minseed=float To use a non-default minimum seed. Default is 1e-3.
   -o, --higherlevel=filename
                       To use a non-default higher level output file name.
   -p, --place, --folder=foldername
                       To use a different common folder for the output files.
                       If this is not provided, the the folder used will be the
                       same as the input folder.
   -r, --relfile, --relationsfile=filename
                       Relations file, with identificators of the higher level
                       in the first column, and identificators of the lower
                       level in the second column.
   -R, --randomise, --randomize
                       A modified version of the relations file is used, where
                       the higher level elements (first column) are replaced by
                       numbers and randomly written in the first column. The
                       numbers range from 1 to the total number of elements.
                       The second column (containing the lowel level elements)
                       remains unchanged.
   -s, --no-steps      Do not print result summary and the steps of every
                       Levenberg-Marquardt iteration.
   -t, --graphtitle=string
                       The graph title (default is
                       "Zij graph for sigma^2 = [variance]").
   -T, --minimalgraphticks
                       It will only show the x secondary line for x = 0, and
                       none for the Y axis (useful for publishing).
   -u, --lowernormw=filename
                       To use a non-default lower level output file name,
                       setting W as weight (default suffix is _lowerNormW).
   -U, --lowernormv=filename
                       To use a non-default lower level output file name,
                       setting V as weight (default suffix is _lowerNormV).
   -v, --var, --varianceseed=double
                       Seed used to start calculating the variance.
                       Default is 0.001.
   -V, --varfile=filename
                       Get the variance value from a text file. It must contain
                       a line (not more than once) with the text
                       "Variance = [double]". This suits the info file from
                       another integration (see -L).
   -W, --graphlinewidth=float
                       Use a non-default value for the sigmoid line width.
                       Default is 1.0.
   -w, --varconf=integer
                       Get the confidence limits of the variance using n
                       by performimg n simultaions
   -y, --varconfpercent=float
                       Get the higher and lower limits to calculate the limits
                       of the variance (see -w). Default is 0.05.
   -z, --outstats=filename
                       To use a non-default stats file name.
   --tags=string       To define a tag to distinguish groups to perform the
                       integration. The tag can be used by inclusion, such as
                            --tags="mod"
                       or by exclusion, putting first the "!" symbol, such as
                            --tags="!out"
                       Tags should be included in a third column of the
                       relations file. Note that the tag "!out" for outliers is
                       implicit.
                       Different tags can be combined using logical operators
                       "and" (&), "or" (|), and "not" (!), and parentheses.
                       Some examples:
                            --tags="!out&mod"
                            --tags="!out&(dig0|dig1)"
                            --tags="(!dig0&!dig1)|mod1"
                            --tags="mod1|mod2|mod3"
   --emergencyvariance In the case the maximum iterations are reached (see -m),
                       force the seed variance as emergency variance.
   --xlabel=string     Use the selected string for the X label. Default is
                       "Zij". To remove the label, use --xlabel=" ".
   --ylabel=string     Use the selected string for the Y label. Default is
                       "Rank/N". To remove the label, use --ylabel=" ".

examples (use "sanxot.py" if you are not using the standalone version):

* To calculate the variance starting with a seed = 0.02, using a datafile.txt
and a relationsfile.txt, both in C:\temp:

sanxot -dC:\temp\datafile.txt -rrelationsfile.txt -v0.02
```

```
* To get fast results of an integration forcing a variance = 0.02922:

sanxot -dC:\temp\datafile.txt -rrelationsfile.txt -f -v0.02922

* To get an integration forcing the variance reported in the info file at
C:\data\infofile.txt, and saving the resulting graph in C:\data\ instead
of C:\temp\:

sanxot   -dC:\temp\datafile.txt   -rrelationsfile.txt   -f   -VC:\data\infofile.txt   -
GC:\data\graphFile.png
```

2.10 SanXoTGauss

```
SanXoTgauss v0.22 is a program made in the Jesus Vazquez Cardiovascular
Proteomics Lab at Centro Nacional de Investigaciones Cardiovasculares, used to
depict the sigmoids of lower level elements compared to their higher levels.
For example: when integrating proteins into categories, the outStats from the
protein confluence will be used along with a list of categories to compare the
sigmoid representing the proteins of each category.

SanXoTgauss needs three input files:

     * a stats file, the outStats file from SanXoT (using the -z command)
     * a higher level list to graph (using the -c command)
     * a relations file (using -r command)

And delivers four output file:

     * the data file used to draw gaussians (default suffix: "_outSigmoids")
     * an extra table with a different arrangement of the previous one (default
          suffix: "_extraTable")
     * the graph (default suffix: "_outGraph")
     * the log file (default suffix: "_logFile")

Usage: sanxotgauss.py -z[stats file] -r[relations file] -c[higher level list file]
[OPTIONS]

   -h, --help          Display this help and exit.
   -a, --analysis=string
                       Use a prefix for the output files. If this is not
                       provided, then the prefix will be garnered from the
                       stats file.
   -c, --list=filename The text file containing the higher level elements whose
                       sigmoids we want to graph. If the first element is not
                       taken, it might help saving the file with ANSI format.
   -d, --graphdpi=integer
                       Set a non-default graph size in dpi (dots per inch).
                       Default is 300 dpi.
   -g, --no-graph      Do not show the sigmoids graph (the file will be saved
                       in any case).
   -G, --outgraph=filename
                       To use a non-default name for the graph file.
   -k, --no-legend     Do not show the legend in the graph (useful when the
                       legend covers the graph, in which case we might want to
                       save it twice: one with legend, and again without it).
   -l, --graphlimits=integer
                       To set the +- limits of the Zij graph (default is 6). If
                       you want the limits to be between the minimum and
                       maximum values, you can use -l.
   -L, --logfile=filename
                       To use a non-default name for the log file.
   -o, --outputfile=filename
                       To use a non-default file name for the sigmoid table.
   -p, --place, --folder=foldername
                       To use a different common folder for the output files.
                       If this is not provided, the the folder used will be the
                       same as the stats file folder.
   -r, --relfile, --relationsfile=filename
                       Relations file, with identificators of the higher level
                       in the first column, and identificators of the lower
                       level in the second column.
   -s, --graphfontsize=float
                       Use a non-default value for legend font size. Default
                       is 8.
```

```
   -t, --graphtitle=string
                       The graph title (default is "Z plot").
   -T, --minimalgraphticks
                       It will only show the x secondary line for x = 0, and
                       none for the Y axis (useful for publishing).
   -W, --graphlinewidth=float
                       Use a non-default value for the sigmoid line width.
                       Default is 1.0.
   -x, --extratable=filename
                       To use a non-default file name for the extra table.
   -z, --outstats=filename
                       The outStats file from a SanXoT integration.
   -Z, --labelfontsize=float
                       The font size used for the labels in the X and Y axes.
                       (Default is 12.)
   --xlabel=string     Use the selected string for the X label. Default is
                       "Zij". To remove the label, use --xlabel=" ".
   --ylabel=string     Use the selected string for the Y label. Default is
                       "Rank/N". To remove the label, use --ylabel=" ".
```

2.11 SanXoTSieve

```
SanXoTSieve v0.12 is a program made in the Jesus Vazquez Cardiovascular
Proteomics Lab at Centro Nacional de Investigaciones Cardiovasculares, used to
perform automatical removal of lower level outliers in an integration
performed using the SanXoT integrator.

SanXoTSieve needs the two input files of a SanXoT integration
(see SanXoT's help): commands -d and -r, respectively.

And the resulting variance of the integration that has been performed:
commands -V (assigned from the info file of the integration.) or -v.

... and delivers two output files:

     * a new relations file (by default suffixed "_tagged"), which is
     identical to the original relations file, but tagging in the third column
     the relations marked as outlier.

     * the log file.

Usage: sanxotsieve.py -d[data file] -r[relations file] -V[info file] [OPTIONS]

   -h, --help          Display basic help and exit.
   -H, --advanced-help Display this help and exit.
   -a, --analysis=string
                       Use a prefix for the output files. If this is not
                       provided, then the prefix will be garnered from the data
                       file.
   -b, --no-verbose    Do not print result summary after executing.
   -d, --datafile=filename
                       Data file with identificators of the lowel level in the
                       first column, measured values (x) in the second column,
                       and weights (v) in the third column.
   -D, --removeduplicateupper
                       When merging data with relations table, remove duplicate
                       higher level elements (not removed by default).
   -f, --fdrlimit=float
                       Use an FDR limit different than 0.01 (1%).
   -L, --infofile=filename
                       To use a non-default name for the log file.
   -n, --newrelfile=filename
                       To use a non-default name for the relations file
                       containing the tagged outliers.
   -o, --outlierrelfile=filename
                       To use a non-default name for the relations responsible
                       of outliers (note that outlier relations are only saved
                       when the --oldway option is active)
   -p, --place, --folder=foldername
                       To use a different common folder for the output files.
                       If this is not provided, the folder used will be the
                       same as the input folder.
   -r, --relfile, --relationsfile=filename
                       Relations file, with identificators of the higher level
                       in the first column, and identificators of the lower
```

```
                    level in the second column.
   -u, --one-to-one    Remove only one outlier per cycle. This is slightly more
                       accurate than the default mode (where the outermost
                       outlier of each category with outliers is removed in
                       each cycle), but usually exacerbatingly slow.
   -v, --var, --varianceseed=double
                       Variance used in the concerning integration.
                       Default is 0.001.
   -V, --varfile=filename
                       Get the variance value from a text file. It must contain
                       a line (not more than once) with the text
                       "Variance = [double]". This suits the info file from a
                       previous integration (see -L in SanXoT).
   --oldway            Do it the old way: instead of tagging, create two
                       separated relation files, with and without outliers.
   --outliertag=string To select a non-default tag for outliers (default: out)
   --tags=string       To define a tag to distinguish groups to perform the
                       integration. The tag can be used by inclusion, such as
                            --tags="mod"
                       or by exclusion, putting first the "!" symbol, such as
                            --tags="!out"
                       Tags should be included in a third column of the
                       relations file. Note that the tag "!out" for outliers is
                       implicit.
                       Different tags can be combined using logical operators
                       "and" (&), "or" (|), and "not" (!), and parentheses.
                       Some examples:
                            --tags="!out&mod"
                            --tags="!out&(dig0|dig1)"
                            --tags="(!dig0&!dig1)|mod1"
                            --tags="mod1|mod2|mod3"
```

2.12 SanXoTSqueezer

```
SanXoTsqueezer v0.10 is a program made in the Jesus Vazquez Cardiovascular
Proteomics Lab at Centro Nacional de Investigaciones Cardiovasculares, used
mainly to find, while using the Systems Biology triangle, which categories
contain a determined number of proteins that are changing more than an FDR
set by the user.

SanXoTsqueezer needs two input files:

     * a lower level stats file (command -l)
     * a higher (or upper) level stats file (command -u)

And delivers one output file:

     * the list of changing higher level elements (which can be used as
     SanXoTGauss input to depict gaussians of relevant higher level elements).

Usage: sanxotaqueezer.py -l[lower level stats file] -u[upper level stats file] [OPTIONS]

   -h, --help          Display this help and exit.
   -a, --analysis=string
                       Use a prefix for the output files. If this is not
                       provided, then the prefix will be garnered from the
                       stats file.
   -f, --fdr=float     FDR to filter data. Default is 0.05. If -z is used, then
                       the program will filter only by Z.
   -l, --lowerstats=filename
                       The lower level stats input file.
   -L, --logfile=filename
                       To use a non-default name for the log file.
   -n, --minelements=integer
                       The minimum number of lower level elements that a higher
                       level element must include in the stats file. Default
                       is 2 (the minimum).
   -N, --maxelements=integer
                       The maximum number of lower level elements that a higher
                       level element must include in the stats file. Default
                       is 1e6.
   -o, --outputfile=filename
                       To use a non-default file name for the sigmoid table.
   -p, --place, --folder=foldername
                       To use a different common folder for the output files.
```

```
                    If this is not provided, the folder used will be the
                    same as the lower stats file folder.
-u, --upperstats=filename
                    The higher (upper) level stats input file.
-z, --sigmas=float  Filter by Z (the number of sigmas the higher level
                    element is deviating from the average). Note that by
                    using this option, you will prevent the program from
                    filtering by FDR (see -f).
```

Appendix 3: Awarded poster at the 13th Human Proteome World Congress, Madrid 2014